EXAMINING STEM-LOOPS AS A SEQUENCE SIGNAL FOR IDENTIFYING STRUCTURAL RNA GENES

by

Kirt Noël
Post Baccalaureate Diploma, Cytogenetics Technology,
British Columbia Institute of Technology, 1997
B. Sc., Biology, University of British Columbia, 1995

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science
in the School
of
Interactive Arts and Technology

SIMON FRASER UNIVERSITY
Spring 2005

APPROVAL

Name: Kirt Noël

Degree: Master of Applied Science

Title of project: Examining Stem-loops as a Sequence Signal for Identifying Structural RNA Genes

Examining Committee:

Dr. Rob Woodbury, Chair,
Professor, SIAT, SFU

Dr. Kay C. Wiese, Senior Supervisor,
Assistant Professor, Computing Science, SFU

Dr. Peter Unrau, Supervisor,
Assistant Professor,
Molecular Biology and Biochemistry, SFU

Dr. Marek Hatala, Supervisor,
Assistant Professor, SIAT, SFU

Dr. Holger H. Hoos, External Examiner,
Assistant Professor, Computing Science,
University of British Columbia

Date Approved: March 21, 2005

Abstract

This project examines stem-loops as a means to identify structural RNA genes along genomic sequences. To undertake this task, an algorithm was developed to scan genomic sequences for stem-loops similar to those typically found in ribosomal RNA. Each stem-loop is quantified with metrics which measure length and spacing attributes. With the help of annotated genomes, we are able to calculate the mean metric values in the various domains which make up the genome. This includes coding sequences, non-coding DNA, ribosomal RNAs, and transfer RNAs. Subsequently, these values are evaluated for their ability to distinguish structural RNAs from their genomic counterparts. Our results indicate that some stem-loop metrics are capable of identifying ribosomal RNA genes in genomes across a wide range of G+C content levels. These results merit further study into this novel approach.

To my charismatic brother and cherished friend Mayco.

"Success is the ability to go from failure to failure without losing your enthusiasm." –
PRIME MINISTER WINSTON CHURCHILL, *1874 - 1965*

Acknowledgments

First, I would like to thank Dr. Kay Wiese for warmly welcoming me to his lab and helping to launch my career in Bioinformatics. Kay took the time to insure I had all the tools necessary to complete this project. His insight and curiosity have proven valuable. Kay provided the financial support to attend the ISMB 2004 Conference in Glasgow, Scotland and the IEEE CSB 2003 and 2004 conferences at Stanford University in California. I am very grateful for your patience, guidance, support, and the freedom you afforded me.

In pursuit of an interesting research project I approached Dr. Peter Unrau for suggestions. He pointed to the absence of computational tools which are capable of finding RNA genes. Our discussion sparked the beginning of this project. Peter also served on my committee. Thank-you Peter.

Alain Deschênes is a fellow graduate student in the Wiese lab. During the course of this project Alain patiently answered countless questions – both good and bad. Alain's suggestions opened my eyes to many unrealized possibilities. His tips helped me to "get the computer working for me rather than vice versa." In retrospect, Alain's guidance literally shaved months of time off this project and saved me from many needless headaches. Thank-you Alain.

Andrew Hendriks is another fellow graduate student in the Wiese lab. He also made his time and his suggestions freely available to me. Andrew and I spent an entire semester dissecting various Bioinformatics algorithms for a graduate course. The time we spent together pouring over notes and sketching diagrams turned out to be one of the most challenging,

rewarding, and enlightening experiences during the course of my studies. Thank-you Andrew.

I would like to thank Dr. Marek Hatala for serving on my committee and Dr. Holger H. Hoos for serving as my external examiner.

Lastly, I would like to thank Edward Glen, Herbert Tsang, Gordon Pritchard, Robin H. Johnson, and Pat Lougheed for their support, suggestions, and technical assistance.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The number of genomes available on public databases has exploded in recent years [3, 4]. These genomes hold the blueprint for a diversity of gene products. There are a number of software programs currently available to locate or find protein-encoding genes within these genomes [29, 47]. Yet, an effective and efficient software application to locate novel structural RNA genes has been elusive.

This research document introduces and examines an innovative approach for locating RNA genes along a genomic sequence. In the most basic terms, it explores whether a specific type of structural pattern, which is universally found in structural RNA genes, can be exploited to define regions along a sequence where they occur. This pattern or structure is known as a stem-loop.

This document is divided into 8 chapters. Chapter 1 introduces terms and concepts used to describe RNA structures. This helps to shed light on the reasons why an effective RNA gene-finder has been evasive. Chapter 2 surveys recent efforts by researchers to develop an RNA gene-finding algorithm. In this context, stem-loops are introduced as a novel approach to help develop a more effective and efficient RNA gene-finder. Chapter 3 describes in detail the algorithm developed to identify the stem-loop structures along genomic sequences. Subsequently, Chapter 4 describes how stem-loop features are quantified into metrics which can be used for statistical analysis and inference. Chapter 5 examines the differences in the average stem-loop metric values between structural RNAs and their genomic counterparts - protein encoding sequences (CDS) and non-coding DNA (NC). Chapter 6 examines stem-loops metrics for their ability to identify structural RNAs along a genomic sequence. Chapter 7 outlines areas where future efforts should be directed. Lastly, Chapter 8 concludes with a summary of the project and its accomplishments.

1.1 RNA Sequences

Ribonucleic Acid (RNA) and Deoxyribonucleic Acid (DNA) share many similarities [1]. Both are comprised of a sequence of nucleotides[1]. RNA and DNA sequences have a directionality whereby the start is referred to as the 5' terminus and the end is the 3' terminus. The values 5' and 3' refer to carbon atoms on the pentose sugar molecule which helps to link the bases together. The pentose sugar is one element which distinguishes RNA from DNA. In RNA, the pentose sugar is a ribose; in DNA, the pentose sugar is a deoxyribose. The bases found in DNA include: adenine (A), cytosine (C), guanine (G), and thymine (T). In contrast, the bases found in RNA include: adenine (A), cytosine (C), guanine (G), and uracil (U). In both DNA and RNA, it is the base in each nucleotide which captures all the attention since they are responsible for the patterns or behaviours geneticists aim to decipher and explain.

The vast majority of RNA sequences[2] are generated by a process known as transcription. In transcription, DNA acts as the template for creating a complementary RNA sequence. The Central Dogma of Biology published in the 1950's suggests that DNA is transcribed into messenger RNA (mRNA) which is subsequently translated into proteins. Under this model, RNA simply acts as a "messenger" helping to direct the production of proteins. Proteins, it seemed, acted alone as the biological machinery responsible for catalyzing cellular activity. However, since the Central Dogma of Biology was published a tremendous number of discoveries have been made. Importantly, researchers have identified many genes whose RNA transcript is not translated into protein. Instead, these RNA molecules act as biological machinery responsible for undertaking catalytic or biological functions. These gene products are commonly referred to as noncoding RNAs (ncRNA) or functional RNAs.

This research project focuses on structural RNAs which are best described as a subclass of functional RNAs. RNA genes products have different means of achieving their objectives. Structural RNAs, for instance, exploit a specific 3-dimensional structure to accomplish

[1] The term nucleotide encompasses 3 elements - a pentose sugar, a phosphate, and a nitrogen base.

[2] There are a small number of retroviruses whose genome is comprised of RNA rather than DNA. One well known example includes the HIV virus which is responsible for AIDS. This project does not study RNA genomes.

a specific task. In contrast, small functional RNAs, such as microRNAs (miRNA), take advantage of a specific primary sequence to accomplish a given objective.

1.2 A Glimpse into the RNA World

Properties unique to the RNA molecule allow them to form biological machinery with exclusive capabilities [13, 15, 60]. They form base pairs and thereby fold into shapes that allow them to execute specialized catalytic processes. They mimic the structure of nucleic acids to block translation. RNA gene products also interact with proteins to form complex structures.

There are several examples of structural RNAs which are involved in numerous forms of gene expression control, protein-binding, and associated activities. *E. coli* 6S ribosomal RNA (rRNA) and human 7SK small nuclear RNA (snRNA) make-up part of the RNA-protein complexes responsible for sequence signal recognition which initiate and regulate transcription [22, 30]. U2 snRNA forms the core of the spliceosome [18]. The *Xist* and *Air* genes [7, 56] are 16,500 and 100,000 nucleotides long, respectively. *Xist* is involved in X-chromosome inactivation. *Air* is involved in autosomal gene imprinting. Both of these phenomena are a means of controlling gene expression but currently their mechanisms are poorly understood. The *H19* gene is roughly 1700 nucleotides long; it is expressed in a few specific tissues [6, 42]. However, its function is not well defined either. Telomerase has a core which is comprised of RNA; this RNA core acts as the telomere template which interacts with DNA sequences [37, 50]. The architecture and mechanism underlying RNA gene products is a testimony to their ability to form complex biological machinery. The diversity in these selected examples suggests a wave of unknown structural RNA gene products may await discovery.

It is important to distinguish structural RNAs from miRNA whose functionality is sequence specific rather than structure specific [32]. Recent findings suggest that miRNA transcripts measuring roughly 21-25 nucleotides play a role in developmental timing and tissue specification [31, 33, 35, 51]. Thus far, there are 2 types of miRNA: small temporal RNA (stRNA) and small interfering RNA (siRNA). stRNAs control developmental timing by mediating sequence-specific repression of mRNA translation [31]. siRNAs mediate sequence-specific mRNA degradation in RNA interference. "stRNAs control developmental timing by mediating sequence-specific repression of mRNA translation" [31]. stRNAs and siRNAs are transcribed with flanking regions that create a folded-back or stem-loop structure. The working RNA molecule - approximately 22 nucleotides long - is cut from the double-stranded stem structure by a protein named Dicer [5]. miRNAs have been found to occur in clusters and individually along genomic sequences.

The number of functional RNA genes discoveries is escalating rapidly and as a result the number of researchers interested in studying them is also growing. Several functional RNA genes identified early on were stumbled upon unexpectedly. Now researchers are designing experiments specifically to find functional RNA genes [31, 33, 35]. Quests for miRNAs involve isolating and cloning small RNA transcripts from cellular lysate and substantiating their existence with expressed sequence tag (EST) methods. Other experiments rely on the unusually small size of small nucleolar RNAs [24]. Isolating larger structural RNA transcripts requires alternate means since their size does not readily distinguish them from the vast pool of RNA transcripts populating the cellular lysate.

Computer technology, online databases, and numerous genomic sequences have made gene recognition algorithms a feasible means for gene discovery. Decades of scientific experimentation have deciphered characteristic sequence signals (i.e. consensus sequences) found in protein-encoding genes [16, 41, 58]. These characteristics can be exploited by gene-finding algorithms to identify protein-encoding genes along a genomic sequence. By comparison, the characteristics of structural RNA genes have been sparsely defined [49]. The scarcity of experimental data on structural RNA gene structure is partially due to the fact that they tend to conserve secondary structure over primary sequences [59]. This scarcity is also a consequence of their relatively recent and unexpected emergence as functional gene products and the difficulty in applying Crystallography and NMR to RNA molecules [2, 21, 36]. A review of Crystallographic and NMR methods is available in Söll et al. [57].

1.3 RNA Secondary Structures

The bases of an RNA molecule have the ability to form Hydrogen bonds with their complementary counterparts. This akin to the property which allows 2 complementary DNA strands to form a double stranded helix. With respect to a single strand of RNA, this means that RNA has an inherent ability to fold upon itself to form base pairs and thereby create a stable RNA structure [61].

Typically, RNA structures are described in terms of 2 dimensions. Biologists refer to

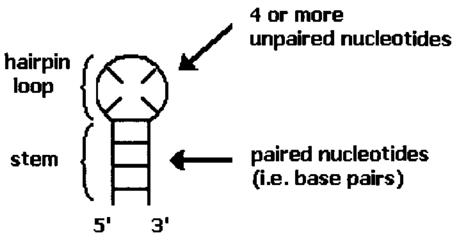


Figure 1.1: The Stem-loop is comprised of 2 components - a hairpin loop and a stem or helix of adjacent base pairs.

2-dimensional structures as secondary structures. Some RNA secondary structures such as ribosomal RNA (rRNA) appear rather complex. Yet these complex structures are simply a culmination of numerous substructures all of which are founded on complementary base pairs. Importantly, there are a number of recurrent patterns seen in these smaller components or substructures. This document focuses on a recurrent elementary RNA secondary structure known as the stem-loop. Stem-loops are universally found in RNA secondary structures.

1.4 The Stem-loop

A stem-loop arises when an RNA sequence folds back on itself (Figure 1.1). The stem-loop can be broken into two components which include a hairpin loop and a stem of base pairs (i.e. a helix) which closes or stabilizes the hairpin loop. The stem-loop is unique in that the base pairs which comprise its helix are only separated by a hairpin loop. When the adjacent base pairs in the stem are interrupted by one or more unpaired bases, the ensuing "bump" is known as a bulge (Figure 1.2) [40]. Similarly, an internal loop forms when mismatched bases are positioned opposite one another in the stem (Figure 1.3) [25, 44, 54].

1.4.1 Stem-loops as Logical Expressions

A more rigorous definition of stem-loops can be formulated with the use of logical expressions. Suppose an RNA sequence consists of n nucleotides. An indexed sequence can be denoted from 5' to 3' as $(0, 1, 2, 3, e, i, k, p, t, n)$ where, $0 < e < i < k < p < t < n$.

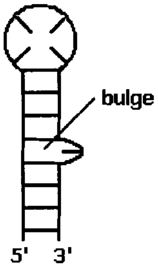


Figure 1.2: A bulge occurs in a stem when it is interrupted an unpaired nucleotide.

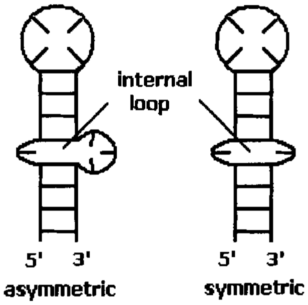


Figure 1.3: An internal loop is characterized by two or more mismatched nucleotides residing in a helix. This figures depicts examples of symmetric and asymmetric internal loops.

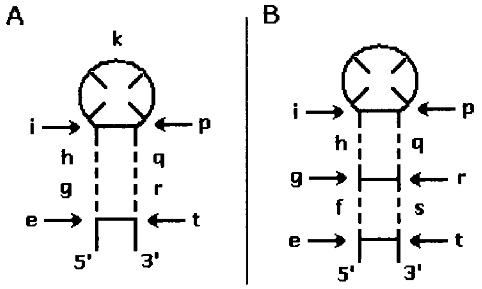


Figure 1.4: Annotated stem-loops. (A) The (i,p) base pair lies nearest the hairpin loop. The (e,t) base pair marks the outer boundary of the stem-loop. (B) A pseudoknot cannot be entirely comprised of nucleotides which lie between the (i,p) and (e,t) base pairs. See text for a detailed description.

Consider a stem-loop where the base pair at the end of the stem nearest the hairpin loop includes nucleotides i and p; this base pair is denoted as (i, p) (Figure 1.4 A). The base pair, (e, t), is furthest from the hairpin loop; it forms the outer boundary of the stem-loop structure. The unpaired nucleotides in the hairpin loop are denoted k; hence, $\forall k \;\; i < k < p$.

For any stem-loop, the nucleotides which lie between e and i cannot base pair with one another nor can the nucleotides between p and t pair with one another (Figure 1.4 A). This rule can be expressed as follows. Suppose a stem-loop is comprised of the following nucleotide sequence: (e, g, h, i, p, q, r, t). If the hairpin is defined by (i, p) and the stem-loop is bound by (e, t) then $\forall g \forall h \;\neg(g, h) \wedge \;\forall q \forall r \;\neg(q, r)$.

Lastly, a pseudoknot[3] cannot be comprised entirely of nucleotides between base pairs (i, p) and (e, t). Suppose a stem-loop is comprised of the following nucleotide sequence: $(e, f, g, h, i, p, q, r, s, t)$. If $(e, t) \wedge (g, r) \wedge (i, p)$: then $\forall f \forall q \;\neg(f, q) \wedge \forall h \forall s \;\neg(h, s)$. See Figure 1.4 B.

[3] When RNA molecules fold, the stems/helices which stabilize these structures generate loops. Pseudoknots arise when nucleotides within a loop base pair with nucleotides outside of this loop. Consider a sequence where: $h < i < j < k < l$. Suppose (i, k) form a base pair. Nucleotide j now lies within a "loop". A pseudoknot arises when either (h, j) or (j, l) form a base pair.

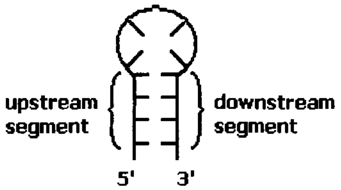


Figure 1.5: The upstream and downstream segments in this stem-loop are labeled. An algorithm generates a list of all the possible upstream segments for a stem-loop with a specific number of base pairs (e.g. 4). Each entry in this list has one or more corresponding downstream segments - i.e. those which are complementary to it. The probabilities of the nucleotides and thereby the probabilities of the upstream and downstream segments can then be used to calculate stem-loop probabilities.

1.4.2 Stem-loop Probabilities

This section discusses stem-loop probabilities and how they are affected by changes to base composition. The probability of a stem-loop is denoted: $P(\textit{stem-loop})$. The probability for each of the possible nucleotides in RNA is denoted: $P(A), P(C), P(G), P(U)$

The most common base pairs found in RNA secondary structures are AU and GC. Less common is the GU base pair. Suppose GU base pairs are *not* observed. This simplifies the task of calculating of $P(\textit{stem-loop})$ since each base has only one complementary partner. Introducing GU base pairs complicates $P(\textit{stem-loop})$ calculations since some nucleotides now have more than complementary counterpart.

An algorithm was devised to calculate $P(\textit{stem-loop})$ where GU base pairs are permitted. This algorithm calculates the probabilities of stem-loops which do not have bulges or internal loops. To accomplish this, the algorithm starts by assembling a list of all the possible upstream segments of a specific size (Figure 1.5). If the number of adjacent base pairs in the stem is 4, there will be 4^4, 256, possible upstream entries in the list. Similarly, if the stem is comprised of 5 adjacent base pairs there will be 4^5 or 1024 upstream segment entries listed. If the upstream segment is comprised of 4 nucleotides and the GU base pair was *not* permitted, there could only be 256 corresponding downstream segments. Importantly, since GU base pairs are permitted, many of the 256 upstream entries will have more than one complementary downstream entry or segment. For instance, upstream segment AAGG can pair with 4 different downstream segments: UUCC, UUCU, UUUC, UUUU. To calculate the probability of finding any stem-loop with 4 adjacent base pairs one needs to add the probabilities for every possible combination - i.e. all the entries in the list. There are certainly more efficient methods to perform these calculations, however, this method suffices for our purposes.

Here is an example of how the probability that a given upstream segment forms a stem-loop is calculated:

$$\begin{aligned}
& \text{Assume}\ldots\ P(A) = P(C) = P(G) = P(U) = 0.25 \\
P(stemloop) &= \sum[P(upstreamsegment) \times P(downstreamsegment)] \\
P(stemloop) &= \Big[(P(AAGG) \times P(UUCC)) + (P(AAGG) \times P(UUCU)) + \\
&\quad (P(AAGG) \times P(UUUC)) + (P(AAGG) \times P(UUUU))\Big] \\
P(stemloop) &= P(AAGG)\Big[P(UUCC) + P(UUCU) + P(UUUC) + P(UUUU)\Big] \\
P(stemloop) &= (0.25^4)\Big[0.25^4 + 0.25^4 + 0.25^4 + 0.25^4\Big] \\
P(stemloop) &= (0.25^4)\Big[(4)(0.25^4)\Big] \\
P(stemloop) &= 6.10 \times 10^{-5}
\end{aligned}$$

In these simplified calculations the number of nucleotides which reside in the hairpin loop is not considered. We will see shortly how this assumption is not detrimental to our $P(stem\text{-}loop)$ calculations.

This algorithm can be run with various nucleotide compositions. However, to maintain a palatable 2-dimensional probability landscape the base compositions where restricted such that $P(G) = P(C)$ and $P(A) = P(U)$. The probability distribution calculated for stem-loops with 4 base pairs over a range of G+C content levels is depicted in Figure 1.6. The lowest probability occurs where $P(G + C) = 0.50$. The highest probability occurs when $P(A + U) = 1.0$ or $P(G + C) = 1.0$.

The shape of the probability distribution function shown in Figure 1.6 may be a bit perplexing at first glance. The next 2 figures are helpful. For a given nucleotide sequence, the probability of finding a GC base-pair increases with increasing G+C content (Figure 1.7). Similarly, the probability of finding an AU base pair increases with increasing A+U content (Figure 1.8). Take note of the inverted x-axis between Figures 1.7 and 1.8.

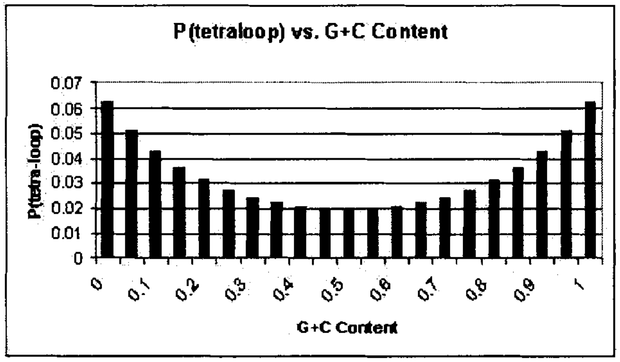


Figure 1.6: Probability of a tetra-loop versus G+C content. The corresponding data is presented in Appendix A.2.

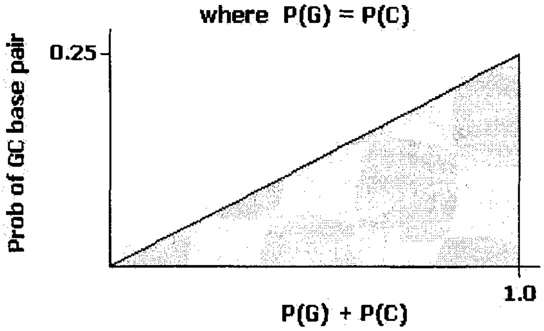


Figure 1.7: Probability of a GC base pair versus G+C content.

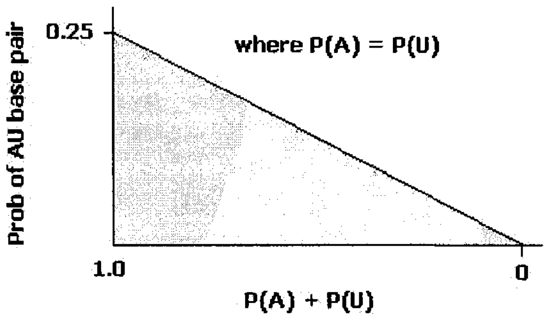


Figure 1.8: Probability of an AU base pair versus A+U content.

Stem Length	*P*(*tetra-loop*)
≥ 4	0.01978
≥ 5	0.007416
≥ 6	0.002781
≥ 7	0.001043
≥ 8	0.0003917
≥ 9	0.0001459

Table 1.1: Probabilities of Tetra-loops with various stem lengths where: $P(A) = P(C) = P(G) = P(U) = 0.25$

In retrospect, Figure 1.6 is simply a combination of Figures 1.7 and 1.8. The advent of GU base pairs does not change the shape of the probability distribution. Instead, it raises the probabilities uniformly.

Another experiment was performed using the aforementioned algorithm to study the effect of stem-loop length on probabilities. First, a fixed G+C content level was selected ($P(A) = P(C) = P(G) = P(U) = 0.25$). The algorithm was repeatedly executed changing only the minimum number of base pairs. The results are presented in Table 1.1. They show that the probability of finding longer stem-loops (i.e. those with more base pairs) is less than the probability of finding shorter stem-loops.

Given *P*(*stem-loop*) and sequence length (n) one can calculate the expected number of stem-loops over a given input sequence:

$$E(stem - loop) = \Big[P(stem - loop) \times (n - 11)\Big]$$

For long sequences...

$$E(stem - loop) \approx P(stem - loop) \times n$$

Each tetra-loop requires at total of 12 nucleotides. Consequently, as one encroaches the downstream terminus there will be 11 consecutive occasions where a tetra-loop could not possibly occur. Hence, the length value must be reduced by 11. For example, in a 100,000 nucleotide sequence with the above composition the expected number of tetra-loops: $E(X) = 1,978$

These simplified probability landscapes have avoided many complex calculations. Nonetheless, these results demonstrate that stem-loops probabilities are affected by base composition. Regions with very high or very low G+C content levels tend to have the most stem-loops. Likewise, regions with very high or very low G+C content tend to have longer stem-loops.

Naturally occurring stem-loops favour a higher G+C content since GC base pairs, which are stabilized by 3 Hydrogen bonds, are more stable than AU base pairs which, in contrast, are stabilized by 2 Hydrogen bonds. Later, the stem-loop search algorithm is described. This algorithm uses search parameters to place constraints on the minimum number of GC base pairs in a stem-loop. This helps to identify stem-loops more in line with what is typically observed in RNA secondary structures.

1.5 RNA Genes Conserve Structure Before Sequence

Like proteins, RNA genes conserve functionality by preserving structure. However unlike proteins, structural RNA genes are not relegated to conserving their primary sequence (ie. the order of their nucleotides) to preserve structure or functionality [49]. Rather, secondary structures can be conserved when interactions between nucleotides in the RNA transcript are preserved. For instance, assume nucleotides in positions 14 to 20 form base pairs with nucleotides 46 to 40 in a given species; these base pairs could be denoted:

$$((14, 46), (15, 45), (16, 44), (17, 43), (18, 42), (19, 41), (20, 40))$$

If the helix formed by these adjacent base pairs is vital to a given RNA structure, we would expect to see nucleotides in these relative positions maintain their complementarity in the same RNA gene found in other highly evolved species. This tendency to preserve structure over primary sequence is described as co-evolution or co-variance [14, 65]. The result, however, is that structural RNA genes are not well-defined by strongly conserved primary sequences. This is why pursuing structural RNA genes computationally is unlike pursuing protein-encoding genes.

Evidently, there are inherent limitations to describing RNA genes and their structures by simply relying on primary sequences. Therefore, researchers have pursued alternate descriptors. One method used by scientists to quantitatively study, analyze, and describe RNA structures relies on the application of thermodynamic models.

1.6 Thermodynamic Models

The stability of RNA secondary structures can be measured using Gibbs Free Energy which is denoted ΔG.

> "Gibbs Free Energy (ΔG) is a thermodynamic property analogous to potential energy. Free Energy is defined so that its change, ΔG, is the negative of the maximum useful work that can be obtained from a reaction at constant temperature and pressure [46]."

A number of thermodynamic models are available to quantify stability in RNA secondary structures [39, 52, 55, 67, 70]. For a given sequence, the most stable RNA structure is the one with the most negative ΔG value. The thermodynamic models for RNA secondary structures assume the sum of the parts equates to the overall ΔG [20, 27, 28]. Consequently, the ΔG of the structural subunits (i.e. helices) contributes additively to the overall ΔG of a complex secondary structure. Generally speaking, the most stable structure is one which uses the optimal combination of substructures.

Interestingly, the lowest ΔG structure is not always the one which is observed in nature. This, however, does not imply that nature favors suboptimal structures. The thermodynamic models generally restrict themselves to secondary (i.e. 2-dimensional) interactions. They do not include tertiary interactions (3-dimensional) or quaternary interactions (between opposing RNA molecules). Therefore, the appearance of "suboptimal" secondary structures in nature could be related to the incomplete state of the thermodynamic models instead of a natural propensity to favour truly suboptimal structures.

Numerous RNA folding algorithms are available to calculate the lowest ΔG structure(s) for a given RNA sequence [11, 19, 38, 43, 48, 62, 69, 68]. These types of applications are commonly used to predict the secondary structure taken-on by a given RNA transcript.

In the next chapter we will see how, researchers have attempted to use ΔG values to identify structural RNA genes along a given genome. The line of reasoning is that genomic segments which code for RNA genes will be marked by particularly low ΔG values since they have evolved to form stable structures. Concurrently, their genomic counterparts are not under the same evolutionary pressures and may therefore have relatively higher ΔG values. If correct, a markedly low ΔG value could thereby act as a statistical signal to identify where RNA genes are located.

1.7 Chapter Review

This chapter introduced RNA molecules. Properties unique to structural RNA genes have been exploited by nature to create a diversity of RNA gene products with highly specialized capabilities. Structural RNA genes preserve structure by maintaining base pair interactions. Proteins, in contrast, conserve primary sequence in an effort to conserve structure and function. As a result, the task of developing a structural RNA gene-finder is unlike that of developing a protein encoding gene-finder.

The next chapter starts with a survey of previous attempts to develop a computational RNA gene-finder. Thereafter, stem-loops are introduced as a possible means to identify where structural RNA genes reside along a given genomic sequence.

Chapter 2

Background

This chapter describes the work done by a number of research groups to develop a computational RNA gene-finder. These approaches follow a similar pattern in that an algorithm divides a genomic sequence into segments and then measures their ΔG values (Figure 2.1). After presenting this research, stem-loops are proposed as a means to help identify structural RNA genes.

2.1 A Brief History of Structural RNA Gene-finders

Le et al. and *Chen et al.* (who incidentally make-up the same group of researchers) published two papers describing how they had used their Z score formula (i.e. Normal Distribution) to assess the significance of RNA transcripts [9, 34]. See Equation 2.1.

$$Z = \frac{\Delta G_{subject} - \overline{\Delta G}_{random}}{s_{random}} \tag{2.1}$$

For a sequence window of fixed length $\Delta G_{subject}$ is the minimum Free Energy is calculated with the use of an RNA folding algorithm. $\overline{\Delta G}_{random}$ is the mean minimum Free Energy for a given population of random sequences with the same length and the same nucleotide composition. s_{random} is the standard deviation of $\overline{\Delta G}_{random}$. To calculate the $\overline{\Delta G}_{random}$ and s_{random} for a nucleotide sequence of the same length and composition as the sequence window. The sequence window was randomly shuffled repeatedly. Each time the ΔG_{random} of this random sequence was calculated using an RNA folding algorithm. Over N random sequences, a mean $\overline{\Delta G}_{random}$ and s_{random} were calculated.

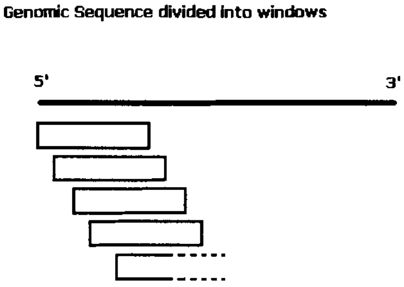


Figure 2.1: Genomic sequence partitioned into windows.

This algorithm has a computational complexity of $O(n^3)$ where n is the sequence length [9]. The authors report promising results using sample sizes, N, ranging from 120 to 2000 and windows (or segments) 30 to 100 nucleotides long. Scanning for structural RNA transcripts involves repeatedly calculating the ΔG for various window sizes and moving along the sequence one base at a time. Monte Carlo simulations are used to assess the statistical significance of the Z scores to identify potentially significant regions and to estimate their optimal (i.e. most probable) size. For this task, the *mFOLD* dynamic programming algorithm was vectorized to run on the Cray X-MP 24 supercomputer.

The reported findings suggest this approach successfully delineates the statistically significant candidate structural RNA transcripts from a pool of random sequences with the same length and base composition. This lends support to the notion that a successful structural RNA gene-finder can be constructed by relying on ΔG values.

In a subsequent article, *Chen et al.* revealed that when their initial method (described above) is carried-out on a supercomputer it is impractical using a window size $>$ 200 nucleotides on an RNA sequence $>$ 1000 nucleotides long [9]. To reduce the computational demands of the algorithm, *Chen et al.* developed a complex formula to estimate mean ΔG_{random} for a nucleotide window equal in size and base composition to the subject sequence. Another formula calculates the optimal window size for the sequence at hand. Whereas the previous method required 150 hours to assess an 800 nucleotide segment, this updated version requires less than 70 seconds while retaining its ability to recognize *structured* RNA transcripts. While the performance has reportedly improved, the updated version fails to address the $O(n^3)$ computational complexity brought about by the RNA folding algorithm used for the $\Delta G_{subject}$ calculation.

A research article by *Rivas and Eddy* disputes the findings published by *Le et al.* and *Chen et al* [9, 34, 49]. It suggests the sequences chosen for analysis misrepresent the ability of the aforementioned ΔG based gene-finder. Rivas and Eddy initially set out to implement a structural RNA gene-finder based on a probabilistic model. Yet, their quest led them to implement two additional models. The following paragraphs explain why.

Unlike the previous group, Rivas and Eddy developed an algorithm to scan sequences using a Stochastic (probabilistic) Context-Free Grammar (SCFG) for RNA transcripts with significant RNA secondary structure. The SCFG model was implemented with a training set consisting of tRNA and rRNA genes to generate a structural RNA gene structure model. They applied an expected log-odds scoring scheme which compares the likelihood that a given sequence has been generated by the structural RNA model or a null model (i.e. something that is not a structural RNA gene) [49].

This SCFG algorithm was executed on a variety of species. They observed that as the genomic A+T composition decreased (i.e. G+C content increased) the log-odds score (i.e. the signal) for tRNA genes diminished. To further examine their suspicions, several species were studied: *M. jannaschii* (30% G+C rich), *C. elegans* (34% G+C rich), *S. cerevisiae* (39% G+C rich), and *E. coli* (50% G+C rich) [49]. Their observations confirmed that their SCFG model was less effective in more G+C rich genomes. This led them to assemble a base composition algorithm to examine how it compared to the SCFG model.

The base composition model simply searched for deviations from the expected base composition in a given species. No structural features are considered in this approach. A log-odds score is calculated for each scanned window. In theory, regions with high G+C content are better candidates for structural RNA genes since GC base pairs are more stable and thereby better suited than AT base pairs to stabilize RNA secondary structures. Interestingly, the base composition model returned the same hits as the SCFG model for the same sequence.

Rivas and Eddy suspected their probabilistic model was not adequately recognizing sequence characteristics related to the secondary structures of structural RNA transcripts. To explore this possibility, tRNA gene sequences were shuffled. If their probabilistic algorithm indeed recognized characteristics in structural RNA secondary structures, the shuffled tRNA genes would be overlooked. Surprisingly, the probabilistic model maintained its hits on the shuffled tRNA genes.

Subsequently, *Rivas and Eddy* [49] implemented a replica of the thermodynamic model developed by *Le et al.* [34]. The objective was to compare the results returned by the ΔG approach to the SCFG model by scanning the same sequences. They found that the Z score had a slight bias towards structural RNA transcripts over their shuffled siblings. However, like the SCFG and base composition models, the thermodynamic model also tended to better identify structural RNA genes in A+T rich genomes. Rivas and Eddy suggest that the ΔG values alone are inadequate for finding structural RNA genes.

In concluding, *Rivas and Eddy* suggest that real RNA secondary structures are not "significantly distinguishable" from the predicted structures of random sequences by thermodynamic and statistical means. Hence, structural RNA gene-finding algorithms must incorporate other factors for gene recognition. The next group to join the pursuit for a structural RNA gene-finder seemingly subscribed to this philosophy.

Carter et al. developed a gene-finder, called RNAGENiE, which implemented a machine learning approach known as a neural network [8]. The neural network was trained specifically for each species being analyzed. The training set included both positive examples (i.e. tRNA, rRNA, and structural RNA genes) and negative examples (i.e. non-coding regions). The protocol divides the chromosome sequence into 80 nucleotide segments or windows with consecutive windows overlapping by 40 nucleotides. The neural network had three input parameters: base composition, sequence motifs (specifically tetra-loop[1] motifs), and ΔG.

Contrary to *Rivas and Eddy*, *Carter et al.* suggest that shuffled structural RNA genes do not qualify as non-coding DNA [8, 49]. Non-coding DNA, they submit, has evolved to take on a deliberate primary sequence. Hence, they constructed a training set of non-coding DNA by removing all protein-coding and structural RNA-coding genes the bacterial genomes they studied. They further removed 50 nucleotides flanking both the 5' and 3' ends of these genes in hope of capturing all the pertinent control elements. The authors acknowledge that the non-coding training set could be contaminated with unknown structural RNA genes.

Carter et al. report the average ΔG in structural RNAs and non-coding DNA regions for several genomes. In structural RNA sequence windows, *E.coli* averaged $-2.70\pm0.52\frac{kcal}{80nt.}$ and *M.jannaschii* averaged $-3.68\pm0.72\frac{kcal}{80nt.}$. In non-coding sequence windows, *E.coli* averaged $-2.06\pm0.85\frac{kcal}{80nt.}$ and *M.jannaschii* averaged $-1.34\pm0.66\frac{kcal}{80nt.}$. These findings suggest that

[1] A tetra-loop is a hairpin loop comprised of 4 nucleotides.

ΔG can help to distinguish structural RNA genes[2]. However, *Carter et al.* do not report the ΔG averages for protein-encoding genes. This would have been interesting given the vast majority (i.e. roughly $\geq 90\%$) of bacterial genomes are comprised of protein-encoding DNA. The report later suggests that the neural network relies more heavily on signals other than ΔG to distinguish coding sequences from structural RNAs.

The performance of the RNAGENiE algorithm was evaluated using a correlation coefficient, Q^α ("average of the percentage of correctly predicted positive and percentage of correctly predicted negative windows") [8]. This provided a means to study the effect of modifications to the algorithm. The average Q^α attained for various bacterial organisms was generally in the high 80s to the low 90s. The hyperthermophilic archaeal bacteria obtained the highest reported Q^α, 99.6%. This exceptional accuracy is related to the high G+C content in structural RNA genes in these A+T-rich organisms. Importantly, Q^α only measures the ability to identify known structural RNA genes and not novel structural RNA genes.

Remarkably, RNAGENiE identified a number of structural RNA genes not included in the training set. Publication and database searches revealed that some of these "novel" structural RNA genes were previously identified in unrelated laboratory experiments. Notably, the vast majority of these novel genes were not completely recognized. Rather the algorithm predicted 1 or 2 of the several windows spanned by these novel genes.

The results strongly suggest that RNAGENiE outperforms the previous structural RNA gene-finders. The improvements can be accredited to several factors. The neural network was carefully trained so that the ΔG parameter was able to better distinguish between structural RNA segments and non-coding DNA. The inclusion of base composition and motif sequence recognition were specifically shown to further improve performance. Also, the diverse training set helped the neural network to distinguish the protein coding sequences from the RNA genes.

Unfortunately, the computational complexity of this approach are not reported. The performance hurdles presumably reside in 2 arenas - the BIOPROP neural network and the Vienna RNA package used for ΔG calculations.

[2] *Carter et al.* report the ΔG results in units of $\frac{kcal}{80nt}$. It is presumed that this is an abbreviation for $\frac{kcal/mol}{80nt}$.

2.2 A Novel Stem-loop Centered Approach

Previous attempts to develop a structural RNA gene-finder have largely relied on RNA folding algorithms to calculate the Free Energy (ΔG) of sequence segments or windows along a given genomic sequence [8, 9, 34]. One risk in using this approach is that ΔG, under the current thermodynamic models, typically decreases as the length of the folded segment increases. This occurs regardless of whether the RNA segment folds into a functioning secondary structure. The culprit is simply that longer nucleotides sequences present more opportunities for base pairs to occur when the RNA molecule folds on itself. In addition, the size of these segments has no biological relevance. Instead, the segment size is an input parameter necessitated by and optimized for an RNA folding algorithm. The logic behind this ΔG based approach dictates that regions along a given genomic sequence which code for structural RNAs will foster statistically significant ΔG values which are indicative of a given segments ability to form a remarkably stable RNA secondary structure. However, there is little evidence in support of such an approach [8, 49]. Furthermore, little attention has been paid to the extraneous factors and complications brought about by the anointed segments sizes. Lastly, the $O(n^3)$ computation complexity of RNA folding algorithms is not conducive to scanning large genomic sequences [68].

There should be convincing evidence to justify including new factors or characteristics into an RNA gene-finder. Hence, our goal is to explore the added benefit stem-loop metrics may provide in identifying structural RNA genes.

There are a number of factors which make stem-loops a suitable target or focal point. Pairing rules and sequence directionality (5'→3') obligate RNA sequences to form stem-loops when they fold upon themselves. One can argue that stem-loops are to RNA structures what α-helices and β-sheets are to proteins. Hence, it is speculated that genomic segments which code for structural RNAs may have a higher density of stem-loops than their genomic counterparts. It is further postulated that stem-loops found in regions which code for structural RNA will tend to be longer than those found in their genomic counterparts. Furthermore, searching for stem-loops along a given sequence can be regulated by a set of parameters to favour stem-loops which are characteristically found in structural RNAs. This element of control is lost when the sequence is partitioned into arbitrarily sized windows to calculate the ΔG. Later, a description of how a set of search parameters is devised is presented. Finally, searching for stem-loop structures along a sequence can be accomplished with $O(n)$ time and space complexity where n is the length of the sequence. This will be explained in more detail after the stem-loop search algorithm is presented.

2.3 Chapter Review

This chapter reviewed previous attempts to develop an RNA gene finder. These approaches have tended to rely on ΔG. The results suggest that more capable sequence signals are required to find structural RNA genes in genomic sequences.

Stem-loops have been proposed as a sequence signal. They make an attractive target given that they are universally found in structural RNA gene products. In addition, it is possible to search for stem-loops along a given sequence in $O(n)$ time where n represents the length of the sequence. The next chapter provides a detailed account of how genomic sequences are scanned for stem-loops.

Chapter 3

Methods - Building A Stem-loop Finder

Earlier, two key motivations for studying stem-loops were stated. One, stem-loops may occur in higher frequency along genomic segments which code for structural RNAs than genomic segments which code for proteins or genomic segments which make up noncoding DNA. Two, stem-loops found in genomic segments which code for structural RNAs may be longer than those which are found in genomic segments which code for proteins or genomic segments which are noncoding DNA. If these notions are true, these differences may be useful in identifying where structural RNA genes occur along a given genomic sequence.

To test these suggestions requires a program capable of identifying stem-loops along an RNA sequence. There are several factors and nuances involved in implementing a suitable stem-loop search algorithm. Therefore, the reader is gradually introduced to the stem-loop search algorithm with a basic search algorithm that has a limited search capacity. Thereafter, the more complex logistics and subtleties which allow the algorithm to find more complex stem-loops are presented.

The search for stem-loops is, in large part, aimed towards finding base pairs which occur adjacently. It is these base pairs which comprise the stem or helix in our stem-loops. The permitted base pairs include AU, GC, and GU (Table 3.1). Note that genomic sequences are transcribed into an RNA before embarking on a search for stem-loops.

Structural RNAs tend to favour stem-loops with a relatively high GC base pair composition. Furthermore, the base pairs which make up a stem-loop are commonly interrupted

A-U	C-G	G-U
U-A	G-C	U-G

Table 3.1: Base pairs permitted by the stem-loop finder.

Indexed Nucleotide (nt) Sequence

$\text{nt}_0 - \text{nt}_1 - \text{nt}_2 - \text{nt}_3 - \text{nt}_4 - \text{nt}_5 - \text{nt}_6 - \text{nt}_7 - \text{nt}_8 - \text{nt}_9 - \text{nt}_{10} - \text{nt}_{11} - \text{nt}_{12} - \text{nt}_{13}$

Figure 3.1: An indexed nucleotide sequence.

by mismatched or unpaired nucleotides. Therefore, it becomes important for the search algorithm to accommodate a set of parameters to accommodate these various possibilities. This will be described in more detail later.

3.1 Basic Stem-loop Finding Algorithm

The basic implementation searches for tetra-loops which are characterized by 4 nucleotides in the hairpin loop and at least 4 base pairs in the stem [66]. No mismatched or unpaired nucleotides in the stem are permitted. To pursue these stem-loops, this basic algorithm increments through the RNA sequence and looks for adjacent upstream nucleotides which can base pair to adjacent downstream nucleotides. Stated differently, this basic search algorithm searches for stacks of base pairs separated only by a hairpin loop.

The algorithm starts with the upstream nucleotide (*nt*) positioned at index 3 (nt_3) (Figure 3.1). It then checks whether nt_3 base pairs with the nt_8 (note this leaves room for 4 nucleotides to occupy a hairpin loop). Suppose, the initial upstream nucleotide (nt_3) and the initial downstream nucleotide (nt_8) do not base pair. Consequently, the upstream nucleotide is incremented by 1. As a result, the algorithm checks whether nt_4 can base pair with nt_9. Assume they base pair, this could mean that they form the base pair which resides nearest the hairpin loop. To check for an adjacent base pair, the upstream index is decremented and the downstream index is incremented. The algorithm then checks whether nt_3 and nt_{10} base pair (Figure 3.2). Suppose 4 adjacent base pairs are identified in the segment depicted in Figure 3.2. This stem-loop is stored as an ordered list of the tuples: $\big((4,9),(3,10),(2,11),(1,12)\big)$.

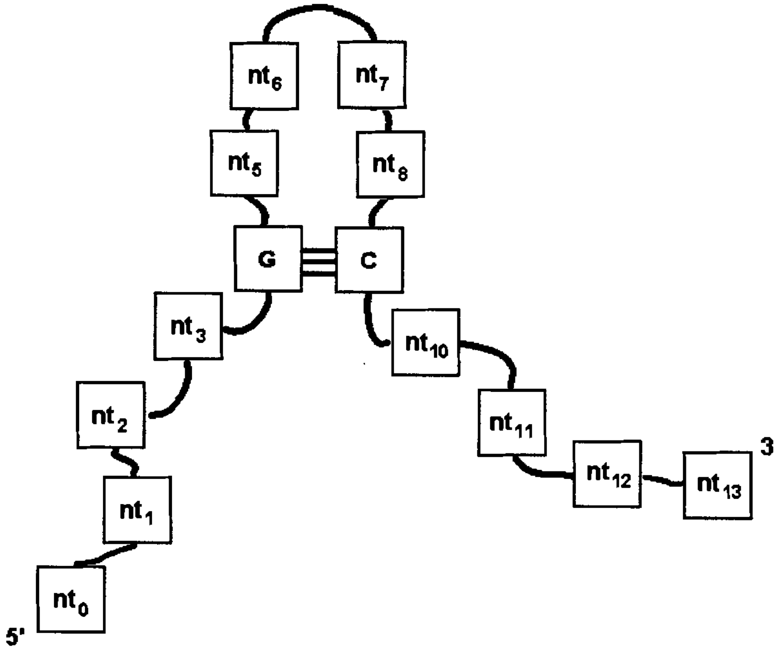


Figure 3.2: The algorithm checks to determine whether complementary nucleotides which form stem-loops can be found. Stem-loops are always validated by starting with the base pair nearest the hairpin loop. In this figure, this involves the nucleotides at index 4 and index 9.

```
function FindTetraLoop(upstreamNucleotide, downstreamNucleotide)
  {
    int n = length of sequence
    tuple bp;
      // stores tuple of indices which denote a single base pair
    pairsInStem
      // stores a list of tuples (i.e. base pairs) which comprise the stem
    int stemLength = 0;
  while( upstreamNucleotide $\geq$ 3 AND downstreamNucleotide leq n-4 AND
  if (sequence[upstreamNucleotide] pairs with sequence[downstreamNucleotide]) )
    {
      bp = (upsteamNucleotide, downstreamNucleotide)
      pairsInStem.append(bp)
        // adds a tuple to the list of detailing the stem
      upsteamNucleotide = upstreamNucleotide - 1
      downstreamNucleotide = downstreamNucleotide+1
      stemLength = stemLength+1
    } end while loop
  if ( stemLength $\geq$ 4 )
    {
      return pairsInStem
    }
  else
    {
      return NULL;
    }
```

Figure 3.3: Pseudocode detailing how tetra-loops are located along an input sequence.

In this simple tetra-loop search, the initial upstream nucleotide and downstream nucleotides are always separated by 4 nucleotides. The subsequent nucleotide pairs in the stem are separated by 6 nucleotides, then 8 nucleotides, and so on (Figure 3.2).

The pseudocode for this basic stem-loop search algorithm is presented in Figure 3.3. In pursuing all the possible tetra-loops along an input sequence, this function is called for all the possible upstream and downstream nucleotides which might form the base pair nearest the hairpin loop. Given a sequence of length n, we can represent the sequence as $\mathrm{nt}_0\mathrm{nt}_1\mathrm{nt}_2\ldots\mathrm{nt}_{n-1}$. Since the minimum stem *requires* 4 base pairs, the lowest index for the upstream nucleotide nearest the hairpin loop is 3. The ordered set depicting the most upstream stem-loop is as follows: $\big((3,8),(2,9),(1,10),(0,11)\big)$. Note, the index values start at 0 and end at $n-1$ for a sequence of length n. Suppose the input sequence has a length of $n = 100$. In this case, the last possible downstream stem-loop which is comprised of the minimum 4 base pairs is defined by the following ordered set of tuples: $\big((91,96),(90,97),(89,98),(88,99)\big)$.

3.2 Partial Validation with Random Sequences

To evaluate the correctness of this simple algorithm, it was implemented and tested on random RNA sequences. One can reasonably conclude that this algorithm functions correctly when it scans random nucleotide sequences and produces results inline with the expected mean given in Table 1.1. To undertake such an evaluation, hundreds of random sequences 100,000 nucleotides long were generated using a function similar to the pseudocode shown in Figure 3.4.

The `RandomlySelectNucleotide(a,c,g,u)` function selects one of the 4 nucleotides at random with each iteration (Figure 3.4). The random sequences generated have a base composition such that: $P(A) = P(C) = P(G) = P(U) = 0.25$

The tetra-loop algorithm was executed on 360 separate random nucleotide sequences each 100,000 nts long. The average number of tetra-loops identified in all of these runs was 1981.59 ± 6.80 (Appendix A.1).

Using the probability from Table 1.1, the expected number of stem-loops is 1977.54 ± 1.96. Recall the probability calculation did not account for the distance between the upstream and downstream segments; this might in part explain the deviation. Nonetheless, the P-value is 0.2676. The fact that these results are statistically close suggests this simple algorithm

```
function GenerateRandomSequence( int length )
    {
        for integer from 1 to length
            {
                randomNucleotide = RandomlySelectNucleotide(A,C,G,U);
                WriteToFile(randomNucleotide);
            }
        return;
    }
```

Figure 3.4: Pseudocode briefly detailing how random sequences are generated.

functions as intended. This is important since this basic algorithm forms the foundation upon which the following more complex stem-loop search algorithm is built.

3.3 Analysis of rRNA Stem-loop Characteristics

Given this algorithm identifies tetra-loops correctly, the next step is to increase its search capacity. This includes the ability to find larger stem-loops and to recognize stem-loops with bulges and/or internal loops. First, there are several questions to address. What is the largest allowable loop? What is the minimum stem-length qualified to stabilize that loop? How big are the largest allowable bulges and internal loops? Are there minimum GC base pair content requirements?

To find answers to these questions, a number of secondary structures were studied. These structures were determined with RNA secondary structure prediction algorithms based on comparative sequence analysis. They are available on public databases at the Comparative RNA Web Site [1] and Ribosomal RNA Database [2] The structures surveyed include: *Escherichia coli* rRNA (Accession Number: J01695), *Saccharomyces cerevisiae* rRNA (V01335), *Coprinus cinereus* rRNA (M92991), *Methanococcus jannaschii* rRNA (U67517) and *Chlamydomonas reinhardtii* rRNA (M32703).

A number of trends were observed. The smallest hairpin loop is comprised of 3 nucleotides; these structures made-up roughly 10% of all the stem-loops. The largest hairpin

[1] Comparative RNA Web Site: http://www.rna.icmb.utexas.edu

[2] rRNA Secondary Structure Models: http://www.psb.ugent.be/rRNA/secmodel/index.html

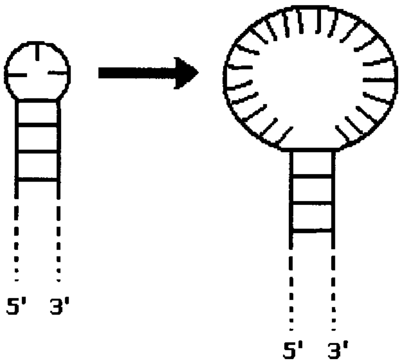


Figure 3.5: In the cursory analysis done on several structures, the size of the hairpin loop ranged from 3 to 20 nucleotides.

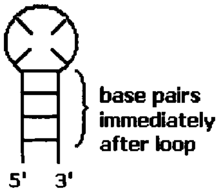


Figure 3.6: Base pairs located immediately after the hairpin loop.

loop observed consisted of 20 nucleotides (Figure 3.5). In all 5 structures analyzed (cited above) there were a total of 188 stem-loops of which only 6 had more than 15 nucleotides in the hairpin loop.

A tetra-loop typically has at least four adjacent base pairs in its stem occurring immediately after the loop closure (Figure 3.6). Anywhere from roughly 20% to 60% of the stem-loops in the structures studied had less than 4 adjacent base pairs immediately closing the hairpin loop. However, virtually all of these stem-loops are further stabilized by base pairs which occur after an internal loop or bulge. See Figures 3.7 and 3.8.

Next, consider the internal loops and bulges (Figure 3.9). Of all the 188 stem-loops observed in the analyzed structures only 3 stem-loops had an internal loop with 7 nucleotides along the longest side. The remaining internal loops were comprised of no more than 6

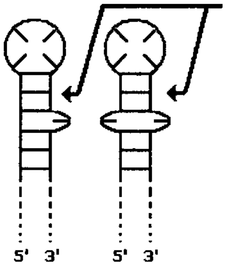


Figure 3.7: Three base pairs in stem before bulge or internal loop.

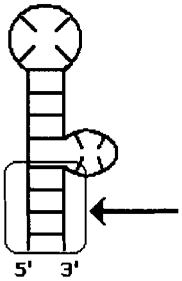


Figure 3.8: Closure - 4 consecutive base pairs distal to this 4 nucleotide bulge.

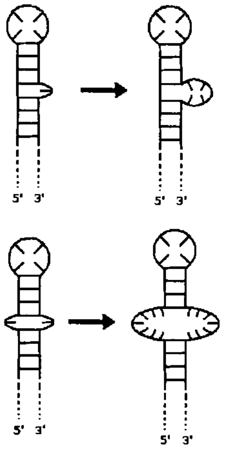


Figure 3.9: Top: Bulges containing 1 to 4 nucleotides. Bottom: Symmetric internal loops containing 1 to 7 nucleotides.

nucleotides. Similarly, there were no instances where a bulge consisted of more than 4 nucleotides. The analysis also suggests that the typical closure stabilizing a bulge is 3 or 4 base pairs long (depicted in Figure 3.8). Similarly, internal loops are typically stabilized by 3 or more adjacent base pairs.

The percentage of GC base pairs in the stem is typically 30–40% or more.

These observations were used to formulate a set of search parameters for our stem-loop search algorithm. A summary of these parameters is presented in Table 3.2.

Presumeably, there are valid objections to the parameters cited in Table 3.2. One of the primary reasons for chosing a fixed set of parameters is to limit the number of variables affecting our results. In the course of this project, several different stem-loop metrics will be tested. Additionally, each of these metrics will be tested on numerous bacterial genomes which have wide ranging G+C content levels. The search parameters have been fixed to make explaining the results more tangible. This does not necessarily deny us the ability to examine the merits of using stem-loop metrics to identify structural RNA genes. Efforts to optimize parameters are more suited to later stages of development.

Parameter	Value
x nucleotides in hairpin loop	$3 \leq x \leq 15$
Min. hairpin loop closure	3 base pairs
Max. bulge	6 nucleotides
Max. internal loop	6 nucleotides
Min. bulge or internal loop closure	3 base pairs
Min. GC Base Pair Content	30%
Max. GU Base Pair Content	34%
Overall min. number of base pairs	4

Table 3.2: Summary of default/initial stem-loop search parameters. Abbreviations: nucleotides (nts), base pairs (bps)

3.4 Extended Stem-loop Finding Algorithm

The algorithm which follows extends the functionality of the basic algorithm described earlier. Recall, it could only find tetra-loops. By adopting the parameters described in the previous section, this "extended" algorithm becomes capable of identifying stem-loops with a range of hairpin loop sizes. It also identifyies stem-loops which are comprised of bulges and internal loops.

The algorithm has been implemented in an object oriented fashion. In adopting this design/implementation approach, the larger goal of finding stem-loops has been broken into smaller tasks or functions. Subdividing the overall goal in this manner adds simplicity to an otherwise complex task. This section describes in detail several of the key functions used to find stem-loops.

The search is coordinated by a function called `FindStems`. The base pairs which occur immediately adjacent to the hairpin loop are referred to as the stem-root (Figure 3.10). As we saw earlier, this is the first section of each stem-loop that is pursued. Hence, the search starts by calling the `FindStemRoot` function. It receives an integer `upstr_nt` corresponding to the index of the first nucleotide in the potential stem-loops upstream segment. Refer to Figures 3.11 and 3.12.

As mentioned earlier, the smallest hairpin loop is a tetra-loop (Figure 3.12). Consequently, the index of the first nucleotide in the downstream segment to check for base pairing is always 5 indices greater. See Figure 3.13. If this pair is complementary, the

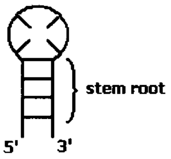
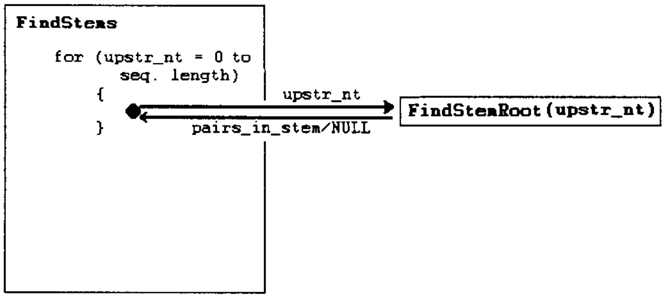


Figure 3.10: The stem-root refers to the base pairs adjacent to the hairpin loop.



```
for (upstr_nt = 0 to
     seq. length)
  {
  }
```

Figure 3.11: Preliminary function diagram.

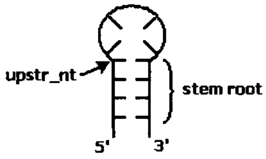


Figure 3.12: First nucleotide in a potential stem-loops uptream segment.

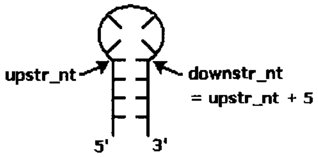


Figure 3.13: First nucleotide in a potential stem-loops downstream segment.

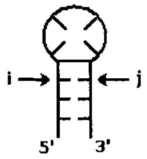


Figure 3.14: After finding a base pair which could mark the beginning of a stem-loop, the indices for the upstream and downstream nucleotides are decremented and incremented by 1, respectively. Using these updated indices the next pair of nucleotides is checked for complementarity.

indices are incremented/decremented by 1 (Figure 3.14).

```
i = upstr_nt - 1
j = downstr_nt + 1
```

This process of decrementing/incrementing the upstream and downstream indices and checking whether the corresponding nucleotides are complementary continues until a pair that is *not* complementary is found. The search parameters dictate at least 3 consecutive base pairs are present in the stem-root. Hence, only when 3 or more base pairs are found does the `FindStemRoot` function return the pairs in the root (`pairs_in_stem`) to the `FindStems` function (Figure 3.11). The `pairs_in_stem` variable might hold something like this:

$$\Big((200, 205), (199, 206), (198, 207), (197, 208)\Big)$$

When less than 3 base pairs are found in the *stem-root* the variable `i` is reset to `upstr_nt` and `j` is set to `upstr_nt + 6`. Now the `FindStemRoot` function looks for a stem-root with a hairpin loop comprised of 5 nucleotides (Figures 3.11 and 3.15). Again, the process of decrementing/incrementing the upstream and downstream indices and checking whether the corresponding nucleotides are complementary ensues. The same rules apply - at least 3 base pairs must be present in the stem-root.

If the size of the hairpin loop exceeds 15 nucleotides and a stem-root with 3 or more base pairs has not been found the `FindStemRoot` function returns `NULL` (Figure 3.11).

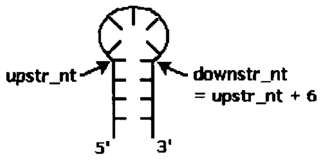


Figure 3.15: First 2 nucleotides to be compared in a potential stem with 5 nucleotides comprising the hairpin loop.

Finding Bulges and Internal Loops

Suppose the `FindStemRoot` function returns `pairs_in_stems`. The next step is to attempt to extend this stem. To find the stem-root the `pairs_in_stems` list was appended until a mismatch was encountered. Therefore, to successfully extend the stem-root around this mismatch the algorithm must accomodate a bulge or an internal loop into the stem. This essentially involves finding a stretch of consecutive base pairs responsible for stabilizing or closing this interruption.

A useful way to describe pursuing bulges and internal loops involves a matrix depicted in Figure 3.16. The upstream segment is anchored along the left side and the downstream segment is anchored along the top side. Note how the consecutive base pairs comprising the stem-root are denoted with cells containing ‘1’. The cell containing a ‘0’ indicates a mismatch. If the goal is to find a string of consecutive base pairs distal to a bulge or an internal loop, the algorithm needs to look for a string of consecutive base pairs (represented by cells labeled ‘1’) occurring somewhere after the mismatch (represented by cells labeled ‘0’). If there is a symmetric internal loop of length 1, then we would expect to find a diagonal of base pairs (or cells labeled ‘1’) interrupted by a mismatch (or cell labeled ’0’) (Figure 3.17). Likewise, if there is a symmetric internal loop of length 2 then we should see 2 mismatches (or cells labeled ‘0’) interrupting the diagonal of base pairs (i.e. cells labeled 1). This is depicted in Figure 3.18.

Let us consider bulges. Suppose, there is a bulge with 1 nucleotide occurring in the upstream segment. The start of the subsequent stretch of consecutive base pairs (those which close the bulge) would start with the cell below the first mismatch (Figure 3.19). Similarly, suppose there is a bulge with 3 nucleotides. Then the subsequent diagonal of base pairs begins 3 cells below the mismatch (Figure 3.20).

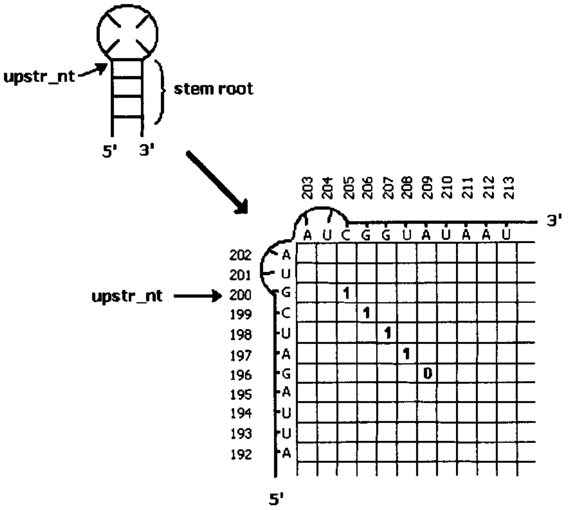


Figure 3.16: Path traversed by consecutive stretch of nucleotides.

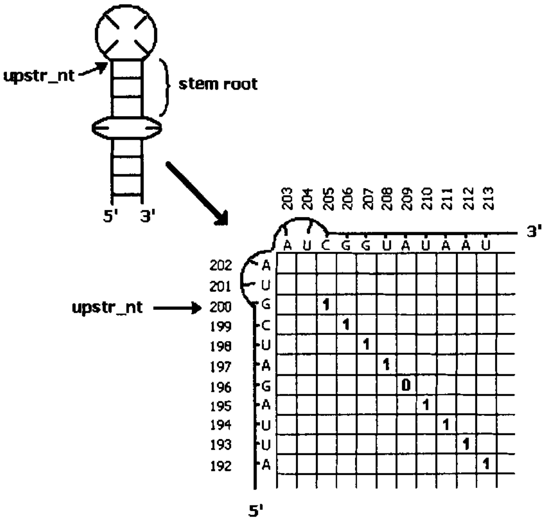
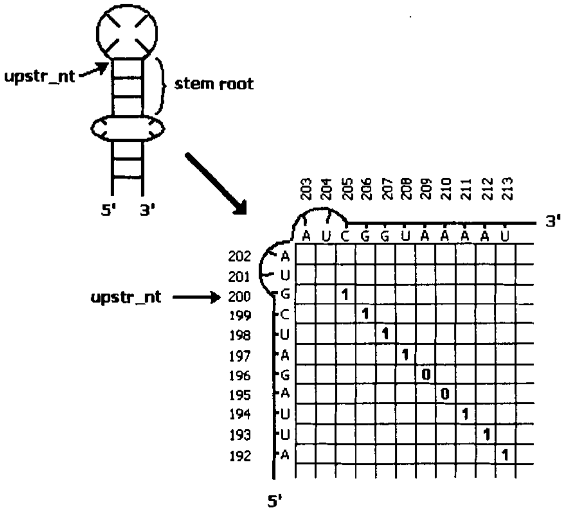
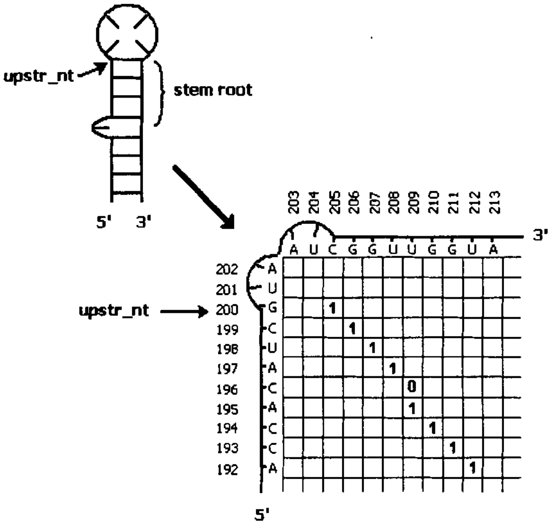


Figure 3.17: Path traveled by symmetric internal loop 1 nucleotide long.



		203	204	205	206	207	208	209	210	211	212	213
		A	U	C	G	G	U	A	A	A	A	U
202	A											
201	U											
200	G			1								
199	C				1							
198	U					1						
197	A						1					
196	G							0				
195	A								0			
194	U									1		
193	U										1	
192	A											1

Figure 3.18: Path traveled by symmetric internal loop 2 nucleotides long.



		203	204	205	206	207	208	209	210	211	212	213
		A	U	C	G	G	U	U	G	G	U	A
202	A											
201	U											
200	G			1								
199	C				1							
198	U					1						
197	A						1					
196	C							0				
195	A							1				
194	C								1			
193	C									1		
192	A										1	

Figure 3.19: Path traveled by upstream bulge 1 nucleotide long.

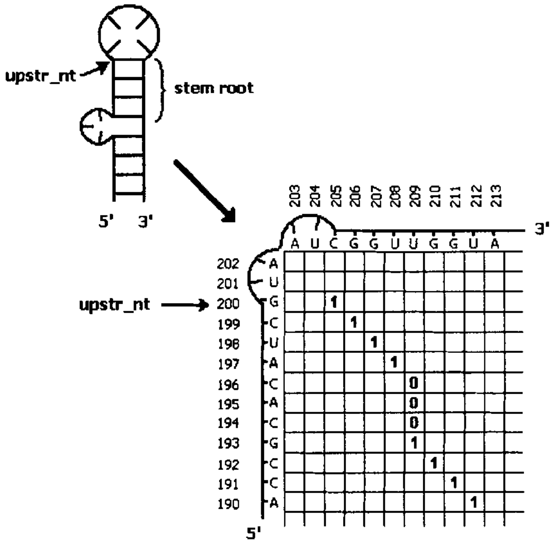


Figure 3.20: Path traveled by upstream bulge 3 nucleotides long.

To reiterate, the diagonal stretch of cells labeled '1' blocks occurring after a mismatch or mismatches represents consecutive base pairs which act to "close" or stabilize bulges and internal loops.

In the previous section, a set of search parameters was outlined (Table 3.2). They limit the maximum number of nucleotides in a bulge or along one side of an internal loop to 6 nucleotides. They also specify that at least 3 base pairs are required to stabilize a bulge or an internal loop. This means that a string of consecutive base pairs must be not less than 3 cells measured diagonally and this diagonal must begin within the 7 cells below or to the right of the mismatch (Figure 3.21).

The examples described above dealt only with bulges and symmetric internal loops. The gray region shown in Figure 3.21 encapsulates all the allowable paths that can be traversed by an asymmetric internal loop. For instance, a diagonal string of base pairs starting one cell to the right and 2 cells below the mismatch (labeled '0') depicts an asymmetric internal loop with 2 nucleotides in its upstream segment and 1 nucleotide in its downstream segment (Figure 3.22).

We now return to the algorithm to see how these tables are useful in describing the

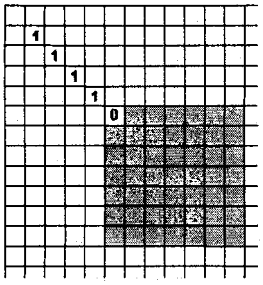


Figure 3.21: This table depicts the search space or the possible starting points for a stretch of consecutive base pairs which close or stabilize an internal loop or bulge.

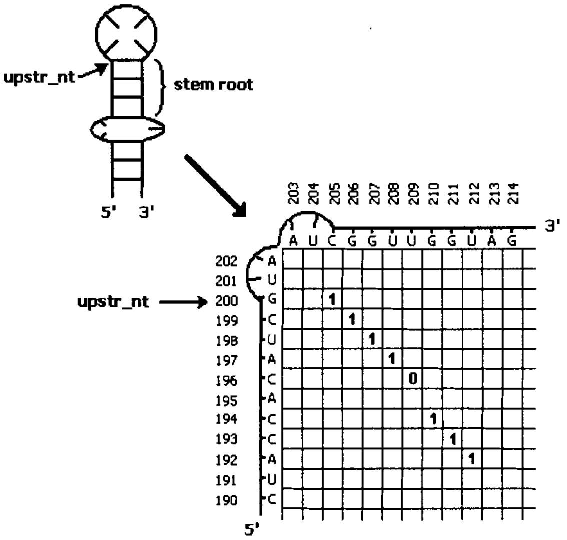


Figure 3.22: Path traveled by asymmetric internal loop.

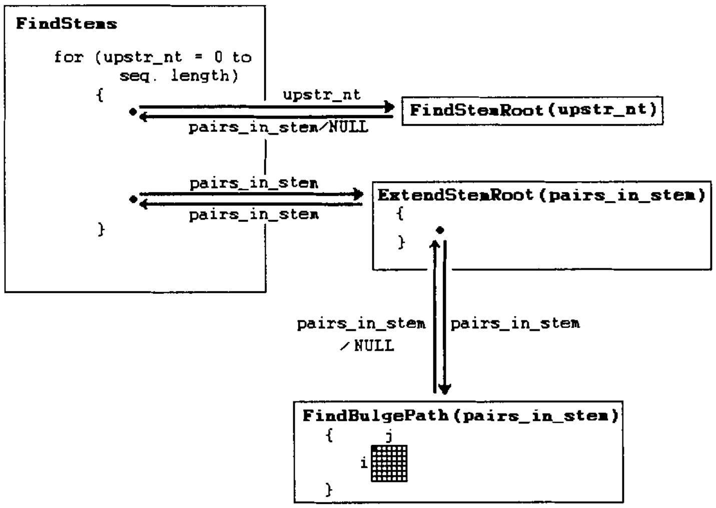


Figure 3.23: Secondary function diagram.

stem-loop search algorithm.

Once the algorithm finds a stem-root it tries to extend the stem either by adopting a bulge or an internal loop. To do this, the `FindStems` function calls the `ExtendStemRoot` function to find the longest possible stem (Figure 3.23).

The `ExtendStemRoot` function calls the `FindBulgePath` function to look for consecutive base pairs which we earlier represented as diagonal paths. When such a path is returned, it is appended to the `pairs_in_stem` variable. `ExtendStemRoot` continues calling the `FindBulgePath` function until it returns `NULL`. In this way, the algorithm seeks out the longest possible stems.

Lastly, two elements need to be added to function diagram (Figure 3.24). For each cell in the block of possible diagonal starting points the `FindBulgePath` function calls the `CheckFor3PlusConsecBPs` function to check whether there are 3 or more consecutive base pairs. The first path with 3 or more consecutive base pairs is not immediately returned. Rather, the first path is stored as a candidate path while the `FindBulgePath` function explores all the (i, j) paths. When a second path is found, the 2 paths are compared by calling the `Select1PathFrom2` function which returns the longest path (ie. the one with the most base pairs).

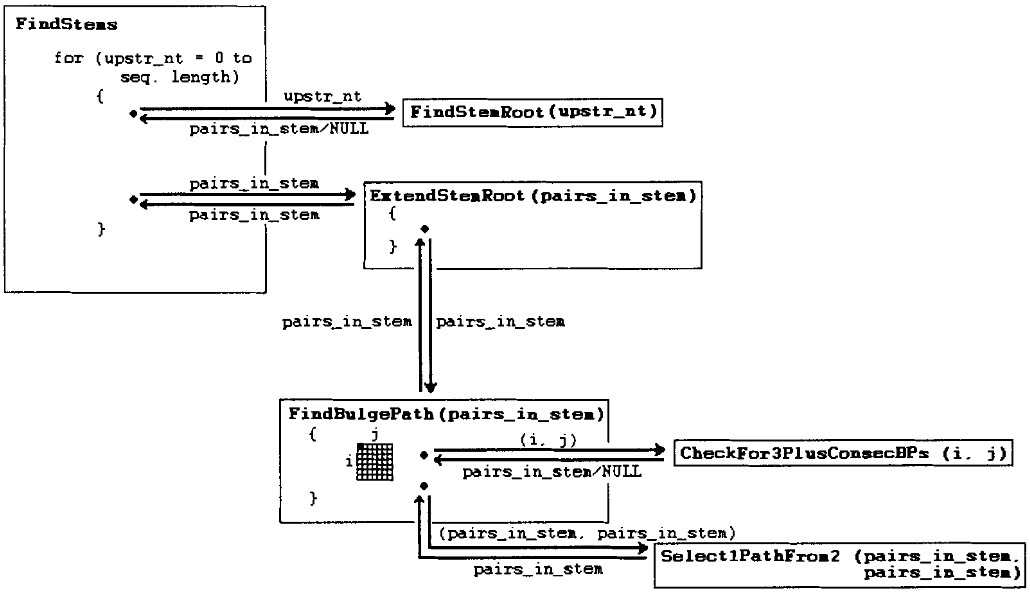


Figure 3.24: Final function diagram.

In summary, at each index along the sequence this stem-loop search algorithm attempts to find the longest possible stack of base pairs with the smallest hairpin loop which satisfies the search parameters.

Earlier, the stem-loop definition presented in Section 1.4.1 explicity stated that stem-loops do not have pseudoknots within their boundaries. It is important to note that this search algorithm does not explicity rule-out this possibility. It is presumed that nucleotides interacting to form a pseudoknot would not reshape or completely destabilize the stem-loops as the algorithm has identified them.

3.5 Computational Complexity

The algorithm scans along a genomic sequence and in doing so it constructs the longest possible stem-loop within the search parameters. It is possible to construct pathological artificial sequences which could lead to a scan time of $O(n^2)$ with a maximum stem-loop size of $\frac{1}{2}n$ (Figures 3.25 and 3.26). However, it seems virtually impossible that such sequences

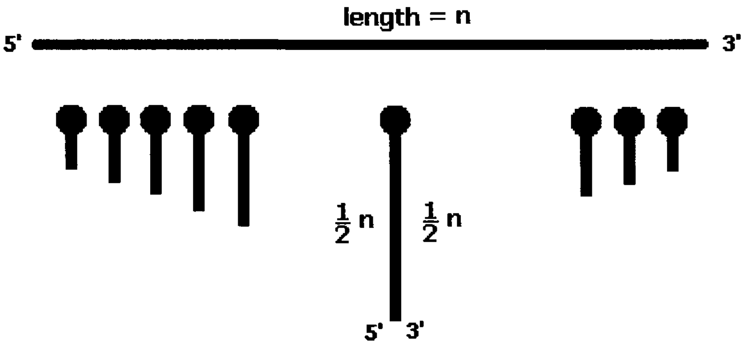


Figure 3.25: In the worst case scenario, a stem-loop is found by the search algorithm at each iteration along the sequence. The longest possible stem-loop would occur at the center of the input sequence. Here approximately $\frac{1}{2}n$ nucleotides in the upstream segment will pair with approximately $\frac{1}{2}n$ nucleotides in the downstream segment.

would occur in nature. For real sequences, the length of the largest possible stem-loop is typically negligible compared to the length of the sequence. For all practical purposes, the average length of a stem-loop can be considered constant and depends more on the G+C content than the length of the input sequence (for large enough sequences). As a result, we can assume that to scan a sequence of length n takes linear or $O(n)$ time. This was also confirmed in practical experiments where the doubling of the length of an input sequence led to roughly doubling the CPU time to scan the sequence assuming that the G+C content of both sequences was the same.

3.6 Programming Language

The stem-loop search algorithm was initially implemented in Python[4] - an interpreted language. Python was selected early on because it allows for rapid prototyping. However, the downside is that its execution is slower than most compiled languages. In Python, the stem-loop search algorithm required approximately 7 minutes to scan through a sequence roughly 1×10^6 nucleotides long (Intel Pentium 4© 2.8 GHz processor, 1.4 GB RAM).

Given that this project would require scanning many sequences - some comprised of over

[4] Python Programming Language: http://www.python.org

$$
\begin{aligned}
s_n &= \text{number of steps for a sequence of length } n\\
s_n &= \text{num. steps in first half} + \text{num. steps in second half}\\
s_n &= \left(\sum_{i=4}^{\frac{n}{2}} i\right) + \left(\sum_{i=1}^{\frac{n}{2}-4} \frac{n}{2} - i\right)\\
s_n &= 2 \times \left(\sum_{i=4}^{\frac{n}{2}} i\right) \quad \text{...simplified since both halfs require the same num. of steps}\\
s_n &= 2 \times \left(\frac{t}{2}(a_1 + a_f)\right) \quad \text{...substitute in a summation series formula}\\
& \ a_1 \text{ is the first value, } a_f \text{ is the final value, } t \text{ is the total number of values.}\\
& \ \text{Since, } a_1 = 4,\ a_f = \frac{n}{2}, \text{ and } t = \left(\frac{n}{2} - 3\right)\ \text{, we have...}\\
s_n &= 2 \times \left[\frac{(\frac{n}{2} - 3)}{2}\left(4 + \frac{n}{2}\right)\right]\\
s_n &= \left(\frac{n}{2} - 3\right)\left(4 + \frac{n}{2}\right) = \frac{4}{2}n + \frac{n^2}{4} - 12 - \frac{3}{2}n\\
s_n &= \frac{n^2}{4} - \frac{1}{2}n - 12\\
s_n &\approx n^2
\end{aligned}
$$

Figure 3.26: In the event that a stem-loop is found at each iteration along the sequence the number of steps (s_n) required to find all these stem-loops is a function of n^2 where n is the length of the sequence.

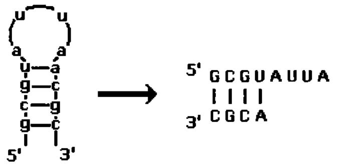


Figure 3.27: Stem-loops are depicted with simple text. The stem-loop on the left is rendered in text format on the right.

5×10^6 nucleotides, these long processing times became an impediment. One of our goals was to design a tool that would give the user a rapid response. Consequently, the algorithm was implemented in the C++ programming language [45]. The performance improvement was substantial. To scan a 1×10^6 nucleotide sequence required less than 6 seconds on the same desktop computer cited above. Simply shifting from Python to C++ appears to have improved time performance by about 70 times. This clearly makes scanning many bacterial genomes much more feasible.

3.7 Displaying Stem-loops in Text Format

The ability to render stem-loops visually is important for inspecting what the stem-loop search algorithm is finding. Consequently, a means to display the stem-loops identified by the search algorithm was devised and implemented.

A simple means to render stem-loops uses 3 lines of text (Figure 3.27). The upstream segment and the nucleotides comprising the hairpin loop are placed on the top line. The bottom line depicts the nucleotides residing in the downstream segment. The vertical bars ('|') located on the centerline signify a base pair between the nucleotides above and below it (Figure 3.28). Mismatched nucleotides do *not* have a vertical bar between them. Unpaired nucleotides are denoted opposite a '*' character.

Recall that the stem-loops are represented in memory as an ordered set of tuples:

$$\big((100,105),(99,106),(98,107),(97,108),(96,109),(92,112),(91,113),(90,114),(89,115)\big)$$

All the information necessary to describe a stem-loop is present in this ordered set. To display the proper nucleotides this algorithm uses the integer values in the tuples to access

```
A G C U G C A U C C G AUUCAUGC
| | | |         | | | | |
U C G A * A A A G G U U
```

```
(89,115) (90,114) (91,113) (92,112) (93,-1) (-94,-111) (-95,-110) (96,109) (97,108) (98,107) (99,106) (100,105)
```

Figure 3.28: A stem-loop depicted using 3 lines of text. Note the mismatched and unpaired nucleotides. The base pair tuples are not included in the output. They are included here for illustration purposes.

the correct locations in the sequence array (e.g. `sequence[100] = A`).

In the interest of memory conservation and efficiency the unpaired or mismatch nucleotides are not stored in the ordered set. Therefore, to include these nucleotides in the stem-loop printout the original ordered set of tuples needs to be modified. However, it is important to clearly convey which tuples represent base pairs and which represent mismatches or unpaired nucleotides. This is accomplished by denoting mismatched nucleotide indices with negative values (e.g. $(-95, -110)$). Similarly, the index of an unpaired nucleotide is paired with a -1 (e.g. $(93, -1)$). This still allows for the proper nucleotides to be retrieved by simply taking the absolute values when *both* integers in the tuple are negative. If only one of the integers is negative, it will be the -1. This signals a mismatch and instructs the algorithm to print a '*' rather than a nucleotide (Figure 3.28). Additionally, one or two negative integers in a given tuple indicates that the center line in the display should *not* include a '|'. The ordered set shown above is processed into the following ordered set to generate the correct text display:

$$\Big((100,105),(99,106),(98,107),(97,108),(96,109),(-95,-110),$$
$$(-94,-111),(-93,-1),(92,112),(91,113),(90,114),(89,115)\Big)$$

Accession Number	Number of Stem-loops Present in Secondary Structure	Identified by Algorithm	Partially Correct	Fully Correct
J01695	32	177	19	9

Table 3.3: This table presents a summary of the findings when the algorithm is executed on a *E. coli* ssu rRNA gene (J01695). Many more stem-loops are identified than are present in the final RNA secondary structure.

3.8 Testing the Search Algorithm on a Single rRNA Gene

The stem-loop finder program was tested on numerous short sequences to insure it was finding the stem-loops as intended. In another evaluation, the stem-loops identified in the *E. coli* ssu rRNA (J01695) gene were compared to the stem-loops depicted in its corresponding RNA secondary structure. The results confirmed previously anticipated output patterns. The algorithm finds many stem-loops; yet it also overlooks others. This commonly occurs when a particular stem-loop fails to meet the search parameters (e.g. $\geq$ 30% GC base pairs). There are 32 stem-loops in the J01695 rRNA secondary structure; the algorithm correctly identified 19 of their locations. This means that the hairpin loop was correctly identified. Identification errors most commonly occur after the algorithm encounters a bulge or an internal loop. At these junctures, the algorithm commonly settles on an incorrect stack of base pairs which closes this bulge or internal loop (Figure 3.21). Recall, the search parameters require 3 consecutive base pairs immediately after an internal loop or a bulge, in some stem-loops there are only 2 adjacent base pairs at these junctures. As a result, the distal segment of several stem-loops were not recorded correctly. The algorithm also misses stem-loops with less than 3 consecutive base pairs in the *stem-root.*

Clearly, the algorithm has not been optimized to identify every stem-loop in a given rRNA structure. However, in this instance, it correctly identify 9 of the 32 stem-loops. Adjustments to the search parameters could help to improve the ability to uncover more "correct" stem-loops. At this juncture, it is too early to determine how such changes would affect the ability to observe differences between structural RNA genes and their genomic counterparts.

Interestingly, the stem-loop search algorithm identifies 177 stem-loops along the J01695 sequence. However, the published secondary structure only depicts 32 stem-loops. This could suggest that every possible stem-loop within a given structural RNA gene is not necessarily present in its final structure. Admittedly, many of these supplementary stem-loops may never have a real possibility of competing - thermodynamically speaking - for a position in the final structure. This, however, does not necessarily preclude the possibility that they may play a role in fostering an environment which favors assembly of the final structure.

The results generated by the stem-loop search algorithm are confined by the search parameters. In addition, the results are also confined by the nature of the search algorithm itself. For instance, the algorithm is designed to take the first "qualified" stem-loop with the smallest hairpin loop. It is important to keep in perspective that the forces at play in determining which stem-loops are present in a final RNA structure are vastly more complex than parameters such as minimum closures, maximum bulge size, or minimum GC base pair content. Hence, it is not surprising that the algorithms findings do not exactly match what is seen in the final secondary structure.

It is postulated that in nature numerous stem-loops in an RNA transcript are competing with one another to play a role in the final structure. Our aim is *not* to precisely define all the stem-loops that would be found in the final RNA secondary structure. Such a task could only be accomplished with an RNA folding algorithm - the problems with that approach were described previously. The questions we hope to answer include the following: Is the average stem-loop longer in regions which code for structural RNAs compared to their genomic counterparts? Do stem-loops occur at a higher frequency in regions which code for structural RNAs compared to their genomic counterparts? The accuracy limitations of the stem-loop search algorithm do not necessarily prevent us from answering these questions.

3.9 Sequence Maps

The motivation for pursuing stem-loops is the notion that stem-loops may occur more frequently in regions which code for structural RNAs relative to the frequency with which they occur in their genomic counterparts. In addition, it is believed that the average stem-loop present in regions which code for structural RNA genes will be longer than the average stem-loop found in their genomic counterparts. To investigate these suggestions requires the ability to compare stem-loops found in the various genomic regions.

Nucleotide Sequence:									
A_0	U_1	C_2	C_3	G_4	A_5	C_6	G_7	U_8	G_9
Corresponding Sequence Map:									
NC_0	NC_1	NC_2	$rRNA_3$	$rRNA_4$	$rRNA_5$	CDS_6	CDS_7	CDS_8	$tRNA_9$

Table 3.4: Annotated sequences are used to create a one-dimensional map. This map depicts the genomic domain a given nucleotide or a stem-loop falls into – ribosomal RNA (rRNA), transfer RNA (tRNA), protein-coding sequence (CDS), non-coding DNA (NC).

Annotated sequences are publicly available at the NCBI[3] website. More specifically, the annotated bacterial sequences used in this study are available at the NCBI Entrez Microbial Genome website[4]. The information conveyed in these annotated genomes is used to generate a one dimensional map for each genome. This sequence map divides the genomic segments into four broad categories - protein coding sequences (CDS), non-coding DNA (NC), ribosomal RNAs (rRNA), and transfer RNAs (tRNA).

For a given annotated genome, two arrays equal in length are created. One holds the nucleotide sequence. The second array serves as a map by storing one of the 4 labels at each index - CDS, NC, rRNA, or tRNA (Table 3.4). This allows the program to determine which domain a specific nucleotide or a stem-loop falls into. For instance, in Table 3.4, the first nucleotide - adenine - falls into a NC region. Similarly, the seventh nucleotide - cytosine - falls into a CDS region.

The information stored in these one dimensional maps is important for comparing stem-loops between opposing regions. Ultimately, this information will help us to determine whether stem-loops can effectively differentiate structural RNA genes from their genomic counterparts.

[3] http://www.ncbi.nlm.nih.gov/

[4] http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html

3.10 Chapter Review

This chapter provided a detailed description of the stem-loop search algorithm. To identify stem-loops similar to those typically observed in RNA secondary structures, a set of search parameters was devised. At each iteration along the input sequence, the algorithm identifies the stem-loop with the smallest hairpin loop and the longest stem which satisfies the search parameters.

Our goal is to determine whether stem-loops can be used to distinguish structural RNAs from their genomic counterparts. To investigate this, the algorithm is used to scan numerous annotated bacterial genomes. From each annotated genome a sequence map is generated. This map divides the genome into 4 broad categories. Subsequently, these maps are used to compare stem-loops identified in structural RNAs (rRNA and tRNA) with those identified in their genomic counterparts (CDS and NC).

The next chapter describes the stem-loop characteristics or metrics which are used to compare stem-loops found in opposing genomic segments. These comparisons involve the use of statistics.

Chapter 4

Methods - Stem-loop Metrics and Statistics

Our goal is to determine whether stem-loops can help to identify structural RNA genes along a given genome. To investigate this requires quantifiable metrics so that statistics can be used to compare stem-loops found in opposing genomic regions. This chapter presents 6 stem-loop metrics. Each metric will be studied to examine its ability to differentiate structural RNAs from their genomic counterparts. This involves comparing the average metric values from structural RNAs (rRNA and tRNA) with the average values measured in their genomic counterparts (CDS and NC).

4.1 Stem-loop Metrics

4.1.1 Number of Base Pairs - *bps*

The *bps* metric is simply the number of base pairs comprising a given stem-loop.

4.1.2 Stem-loop Span - *span*

The *span* metric is measured as the distance in nucleotides from the first upstream nucleotide to the last downstream nucleotide which comprise a stem-loop (Figure 4.1). Stated differently, it is the distance between the 2 nucleotides which makeup the base pair which lies furtherest from the hairpin loop.

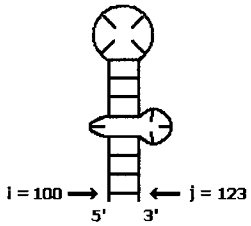


Figure 4.1: The stem-loop *span* is measured as the total distance between the nucleotides which comprise the base pair which is most distal to the hairpin loop.

$$\begin{aligned} span &= j - i + 1 \\ & \text{With reference to Figure 4.1...} \\ span &= 123 - 100 + 1 \\ span &= 24\ nucleotides \end{aligned}$$

4.1.3 Stem-loop Center-point Spacing - *cSpacing*

The center-point for a given stem-loop is the position in the middle of the hairpin loop. The *cSpacing* metric gauges the distance between stem-loops by calculating the average distance in nucleotides from the center-point of a given stem-loop to the center-points of its nearest non-overlapping upstream and downstream neighbours (Figure 4.2). For instance, if the base pair nearest the hairpin loop is $(100, 105)$ the halfway or center-point in the hairpin loop is 102.5. Similarly, if the base pair nearest hairpin loop is $(100, 106)$ the center-point is 103. The center point does not need to correspond to a single nucleotide since it is being used to gauge distance and not to denote which nucleotide lies at the center of a given hairpin loop.

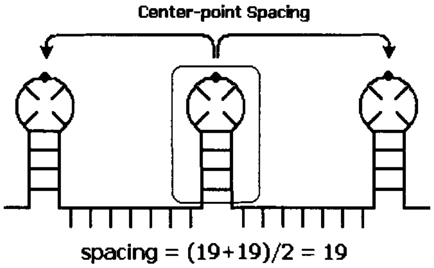


Figure 4.2: The *cSpacing* metric is measured from the center point of each stem-loop indicated by the arrows.

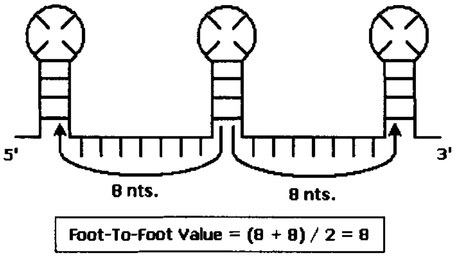


Figure 4.3: The *fSpacing* metric is measured as the average distance in nucleotides from the foot of one stem-loop to the footings of its nearest *non-overlapping* upstream and downstream stem-loops - as indicated by the blue arrows.

4.1.4 Stem-loop Foot Spacing - *fSpacing*

The "foot" of a stem-loop is the base pair most distant to the hairpin loop (Figure 4.3). The *fSpacing* metric is the average distance to the foot of its nearest non-overlapping upstream and downstream stem-loops. The *fSpacing* metric is similar to the *cSpacing* metric, however, it excludes factors related to the size of the stem-loop itself by measuring from the outer boundaries.

The final metrics simply group 2 metrics together. Multiplication, it was postulated, would help to amplify differences that might exist between structural RNAs and their genomic counterparts. In contrast, simply adding 2 metrics together would not significantly amplify differences between the various genomic regions.

4.1.5 Combined Metric - $\left(cSpacing \times bps\right)$

The $\left(cSpacing \times bps\right)$ metric combines the *cSpacing* and the *bps* metrics by multiplying them.

4.1.6 Combined Metric - $\left(fSpacing \times bps\right)$

The $\left(fSpacing \times bps\right)$ metric combines the *fSpacing* and the *bps* metrics by multiplying them.

4.2 Statistics

4.2.1 Mean Metric Values for a Genomic Domain

The sequence maps are used to calculate the mean stem-loop metric values for each of the genomic domains - rRNA, tRNA, CDS, and NC (Equation 4.1).

$$\overline{X} = \frac{\sum_{i=0}^{n} x_i}{n} \qquad \begin{array}{l} x_i \text{ refers to the metric value for stem } i \\ n \text{ is the number of stems found in a region (eg. CDS)} \end{array} \tag{4.1}$$

It is conceivable that a stem-loop will sit atop a boundary between 2 genomic domains. However, it is assumed that any given stem-loop can belong to only 1 genomic domain. The genomic domain for a given stem-loop is determined by rounding the center-point value to the nearest integer or index and looking up that index in the sequence map.

The mean stem-loop metric values are calculated for all 58 genomes in our test set. These results are plotted and compared to study differences and trends. This information helps to determine whether mean stem-loop metric values may be capable of distinguishing structural RNAs from their genomic counterparts. The next section describes a more rigorous statistical examination. Given the average stem-loop metric value in structural RNAs, can they be identified along the sequence without using the sequence map?

4.2.2 Hypothesis Tests Help Identify rRNAs

A statistical inference is commonly referred to as a hypothesis test by statisticians. In this project, hypothesis testing is used to determine whether a given region along the genomic sequence potentially codes for a structural RNA. These statistical inferences are based on the Central Limit Theorem (CLT). According to the CLT, sample means are approximately Normally Distributed about the population mean, μ (see figure in Table 4.1) [12]. This makes it possible to determine whether a given sample mean, $\overline{x}$, may have been obtained from a population with a mean value, μ. This test is commonly based on a 95% confidence interval. Consequently, 2.5% of the extremely low sample means and 2.5% of the extremely high sample means are discarded (Table 4.1). Theoretically, if a 100% confidence interval was desired the thresholds or cut offs would have to extend to negative and positive infinity.

The stem-loop search algorithm identifies the stems in the genome sequence. The sample mean for a given stem-loop metric is calculated over N adjacent stem-loops. This sampling process is repeated over the entire length of the input sequence. For each sample along the sequence, a hypothesis test is performed. If the sample mean lies within the 95% confidence interval of the population mean, μ_{rRNA}, the sample is suspected to have come from an rRNA gene. Note that the classification of each sample is thereby reduced to a true or false answer.

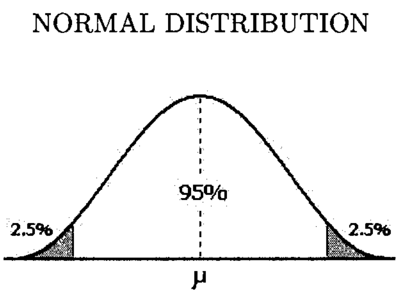
The sample size - N - used in our experiments ranges from approximately 30 to 500 stem-loops. Increases to sample size are accompanied by decreased variance in the sample mean probability distribution (i.e. the Normal Distribution of sample means takes on a more narrow shape) [12].

This project does not apply a strict mathematical method to evaluate false positive rates. It was decided that graphical depiction of the algorithms finding would suffice for our purposes. The next section describes how graphical illustration of the results are generated for each of the input sequences.

4.3 Displaying the Results Obtained on Bacterial Genomes

Once a genome is scanned for stem-loops and the statistical analysis is performed the results are summarized in a figure. To see an example the reader may jump ahead to Figure 6.1 on page 84. This figure is divided into 2 graphs. The top graph shows a map of the annotated genomic sequence with reference to the indexed sequence (i.e. nucleotide location). It also depicts where the "hits" determined by the statistical analysis occur. The ideal result shows a strong correlation between the location of the rRNAs/tRNAs and the "hits". The bottom graph shows local GC content plotted against the same index values.

The results are depicted in 500,000 nucleotide segments. This allows for these graphics



$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

Table 4.1: Normal distributions are symmetrical. The CLT states that sample means approximate a Normal Distribution which is centered over the population mean, μ. A 95% confidence interval encompasses 95% of the possible sample means could arise from this population, μ. The remaining 5% covers the outlying sample means which have a much smaller probability of occurring.

or graphs to be easily presented on paper. In the interest of space, the results reported in this document on any particular genome have been restricted to 1 of these 500,000 nucleotide segments.

4.3.1 Plotting Results with Gnuplot

These graphs along with several others were created using a text based program called Gnuplot. Gnuplot is advantageous when thousands of data items need to be plotted. Loading a text file in Excel or a Statistics Program is generally workable. However, their graphical interfaces crash when one attempts to open a data file several megabytes in size. Gnuplot easily handles large quantities of data. Furthermore, if one is graphing the same data format for numerous different files (ie. sequences) a script can be written to undertake this repetitive task in a hands-free manner. For a detailed description on Gnuplot see their website: http://www.gnuplot.info

4.4 Chapter Review

This chapter described several stem-loop metrics used to study differences between structural RNAs and their genomic counterparts. To distinguish between the various domains our approach relies on stem-loop metrics and their statistics. The aim is to find a set of stem-loop metrics where the average values observed in structural RNAs differ significantly from the average values found in their genomic counterparts. These metrics have been devised to measure the length and spacing attributes of stem-loops.

After scanning the genomes in our training set, average values for each of the stem-loop metrics are established. These average values are used to compare and contrast the various genomic domains - CDS, NC, rRNA, and tRNA. Our final experiments examine whether an average structural RNA metric value is capable of identifying where structural RNAs occur along a given genome without the use of the sequence map.

It is important to test stem-loop metrics on many genomes over a wide range of G+C content levels. This provides a more realistic guage on their effectiveness across a wide diversity of genomes. Consequently, the next chapter describes how the stem-loop metrics are tested on numerous bacterial genomes to study how changes to G+C content affect their performance.

Chapter 5

Results - Average Stem-loop Metric Values

This chapter examines differences between the genomic domains in our set of annotated bacterial genomes. First, the base composition in CDS, NC, rRNA, tRNA regions are compared. This information is useful in understanding the subsequent sections which describe the differences in the stem-loop metrics between the opposing genomic domains.

5.1 Base Composition

The base composition approach studies the frequency of nucleotides along a given genomic sequence. Research has shown that RNA genes tend to be G+C rich (i.e. they have a relatively high makeup of guanine and cytosine nucleotides) [10, 17, 23, 26, 63]. In A+T rich genomes, the average G+C content in RNA genes is considerably higher than the average G+C content found in their genomic counterparts. This disparity can be exploited to uncover where structural RNA genes are located [8, 23, 17, 49, 53, 64]. By simply measuring *local* G+C content (i.e. over a region spanning roughly 100-200 nucleotides) one can identify where RNA genes occur in an A+T rich genome. Importantly however, this base composition approach is considerably less effective when the difference in the *global* G+C content and the structural RNA G+C content decreases.

In the paragraphs which follow, the term "global" is used repeatedly. It refers to the

Genomic Domain	**Mean G+C Content**	**Standard Deviation**
CDS	0.45	0.12
NC	0.39	0.12
rRNA	0.53	0.043
tRNA	0.58	0.048

Table 5.1: Summary statistics for the average G+C content for each genomic domain across the *global* G+C content spectrum. Original data presented in Appendix A.3.

entire genomic sequence. For instance, the *global* G+C content describes the base composition over an entire genomic sequence. Conversely, the term *local* loosely describes a regional characteristic.

The first experiment examines the differences in *local* G+C content between various genomic domains over 58 different bacterial genomes. The results depict interesting trends which emerge in genomes over a wide range of *global* G+C content levels. Figure 5.1 shows the mean *local* G+C content level for each of the genomic domains - CDS, NC, rRNA, and tRNA. Each genome contributes 4 points to the graph - one for each of its genomic domains. It should be clear that the 4 points from any given bacterial genome are aligned vertically since they all arise from the same genome sequence which has only one *global* G+C content value.

There are a couple of noteworthy features in Figure 5.1. The G+C content levels in rRNA and tRNA regions are notably more stable than those found in the CDS and NC regions (Table 5.1). Furthermore, the G+C content levels in CDS and NC strongly correlate with the *global* G+C content levels.

The G+C content level in structural RNAs appears relatively steady at roughly 50-60%. Importantly, when the *global* G+C content reaches 50-60% distinguishing between structural RNAs and their counterparts using differences in base composition becomes infeasible. This conclusion can be made merely by studying Figure 5.1. The plots for rRNA and tRNA intersect with the CDS and NC plots in genomes when the *global* G+C content level reaches 50-60%. This collision illustrates why the base composition method is not effective at finding structural RNAs in G+C rich sequences.

The sections which follow describe the pursuit of a set of stem-loop metrics which are

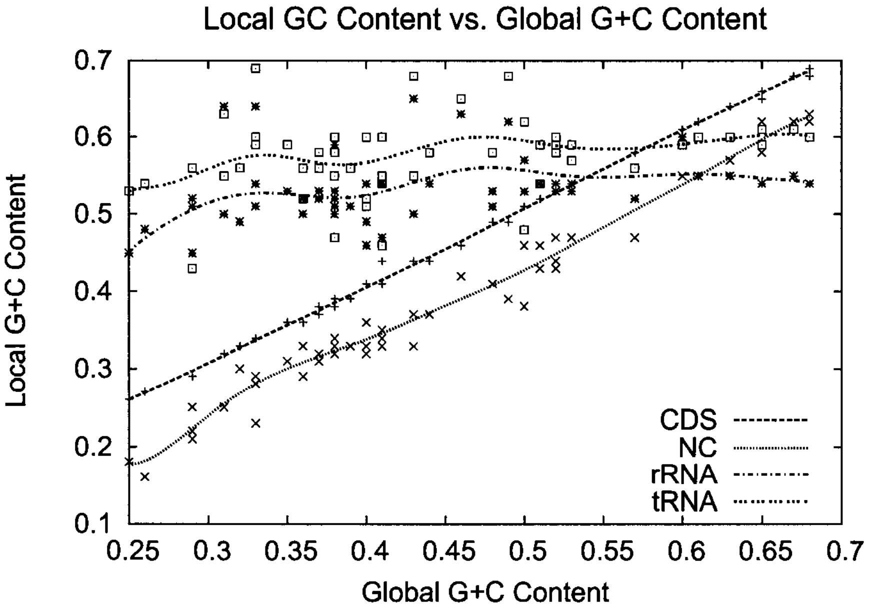


Figure 5.1: Local G+C content vs. Global G+C content. This graph depicts the differences in local G+C between the various genomic domains. As *global* G+C content increases the disparity between the *local* G+C content levels in structural RNAs (rRNA and tRNA) and their counterparts (CDS and NC) diminishes. Where these plots collide, the values for the respective plots are equivalent. The corresponding data is located in Appendix A.3

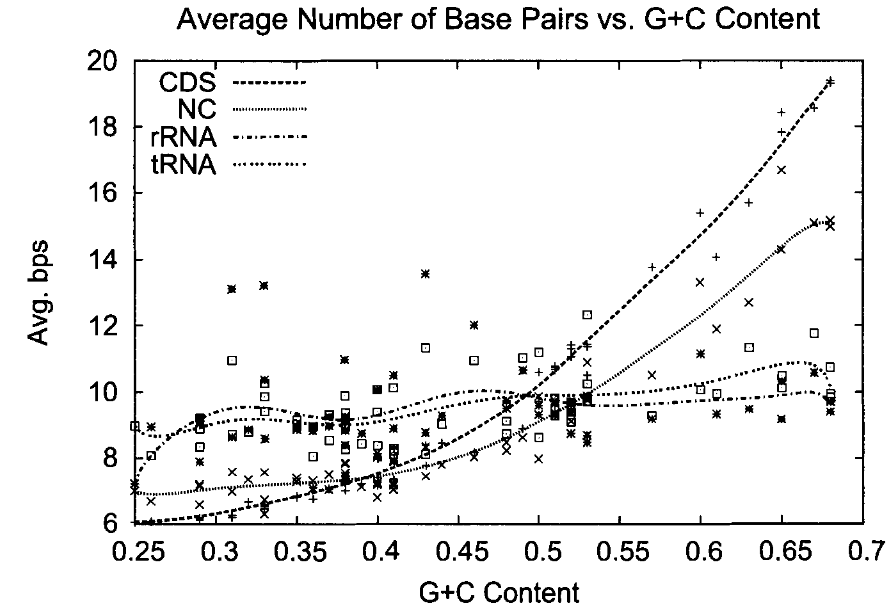


Figure 5.2: Average *bps* found in stems-loops in CDS, NC, rRNA, and tRNA regions. The data is presented in Appendix A.4

capable of identifying regions where structural RNA genes occur. These metrics are tested on many genomes in an effort to evaluate and predict their performance across the *global* G+C content spectrum. The observations in Figure 5.1 convey the value in finding a set of stem-loop metrics where the average values in structural RNAs differ significantly from their genomic counterparts across the entire G+C content spectrum.

5.2 Number of Base Pairs in a Stem-loop - *bps*

The stem-loop base pairs metric, *bps*, is intuitive. It is simply the total number of base pairs in a given stem-loop. Our initial expectations were that the average *bps* in stem-loops which reside in rRNA and tRNA genes would be greater than the average *bps* observed in their genomic counterparts. The results over the same set of 58 bacterial genomes are shown in Figure 5.2.

There are a couple of noteworthy observations. First, the results contradict our initial

Genomic Domain	**Average** *bps*	**Standard Deviation**
CDS	9.64	3.92
NC	8.92	2.63
rRNA	9.44	1.27
tRNA	9.51	1.01

Table 5.2: Summary statistics for the average *bps* in a stem-loop for each genomic domain across the G+C content spectrum. The corresponding data is presented in Appendix A.4.

suspicions. The average length of a stem-loop in rRNA and tRNA regions is *not* consistently greater than the average in CDS and NC domains. Furthermore, there does *not* appear to be a significant difference in the average number of base pairs found in structural rRNA compared to their counterparts. The only exception to this may be in G+C rich genomes. Compared to the previously described base composition approach, the *bps* metric does not appear to provide an advantage in demarcating structural RNAs at any point along the G+C content spectrum. Stated differently, the average *bps* value in rRNA domains compared to the average in its counterparts do not appear to differ significantly in the genomes examined.

Figure 5.1 suggests that increases to *global* G+C content coincide with increases to *local* G+C content in CDS and NC domains. This makes sense given the vast majority of bacterial genomes are comprised of CDS sequences. Recall, the *local* G+C content level in rRNA and tRNA genes are relatively stable by comparison (Table 5.1 and Figure 5.1). A similar trend emerges in Figure 5.2. As the *global* G+C content increases from one genome to the next, the average *bps* in the CDS and NC domains also increases. These results suggest that the increase in average *bps* in CDS and NC regions is strongly related to the concurrent increase in *local* G+C content level in these domains. Likewise, the relatively stable *bps* valuations in rRNA and tRNA domains is presumably related to the relatively stable G+C content levels in these domains (Tables 5.1 and 5.2).

Recall the stem-loop search parameters; they require a minimum 30% of the base pairs are GC base pairs. The higher the G+C content the more likely one will find longer stems with a sufficient number of GC base pairs (Figure 1.6 on page 10). Conversely, in AT rich genomes, the longest possible stems will tend to have more AU base pairs. However, the longest stems in AT rich genomes are less likely to meet the minimum GC base pair

NC_001318 *bps* metric	CDS	6.17 ± 2.21
	NC	6.58 ± 3.06
	rRNA	7.88 ± 3.52
	tRNA	8.25 ± 3.99

Table 5.3: Mean values for *bps* metric in *Borrelia burgdorferi*, NC_001318, which has a global G+C content of 29%.

requirement. Given the stem-loop search algorithm only continues to extend a stem as long as the minimum parameters are met, the average *bps* value in AT rich genomes is bound to be lower.

5.2.1 Probability Distribution - *bps*

For a given sequence, a probability distribution can be created for the *bps* metric. In the interest of space, the distribution of only 1 genome - *Borrelia burgdorferi* - will be presented here (Figure 5.3). The 4 genomic domains are plotted separately. The shape of the distributions among the various domains appears to follow a similar pattern. The mean values occur at slightly different positions (Table 5.3 and Figure 5.3). The relatively high standard deviations in these mean values is attributable to the shape of their probability distributions.

An analysis of the results obtained for the *bps* metric in randomized or shuffled genomic sequences is presented in Appendix B.

5.3 Stem-loop Span - *span*

The span is measured as the distance from the first upstream nucleotide to the last downstream nucleotide comprising the stem-loop structure (Section 4.1.2). Using the same set of 58 genomes, a graph was generated by plotting the average *span* versus the *global* G+C content is remarkably similar to the earlier graph depicting the *bps* metric (compare Figures 5.2 and 5.4). The similarity in the graphs corresponds to their likeness; the *bps* and *span* metrics both relate to the length of the stem-loop. The difference being that the *bps* metric does not account for the unpaired nucleotides which reside in the hairpin loop, the bulges,

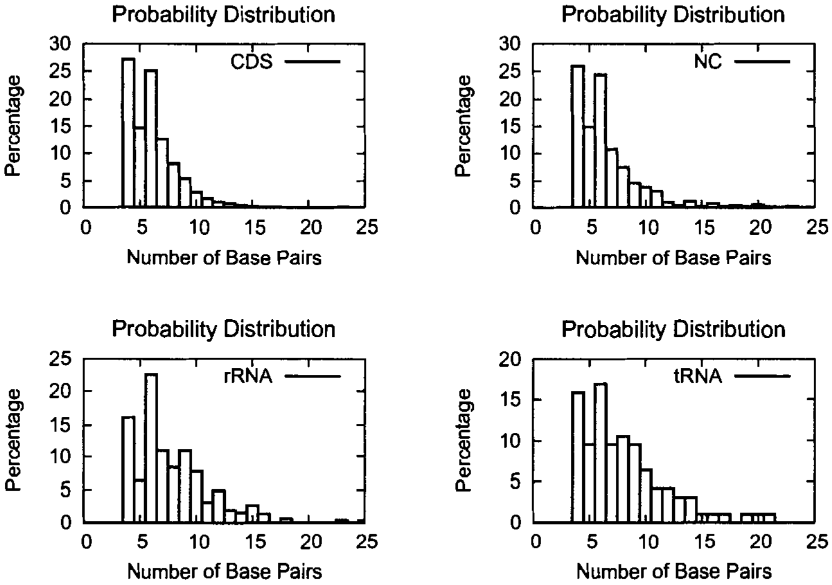


Figure 5.3: Probability distribution of the *bps* metric in *Borrelia burgdorferi*, NC_001318.

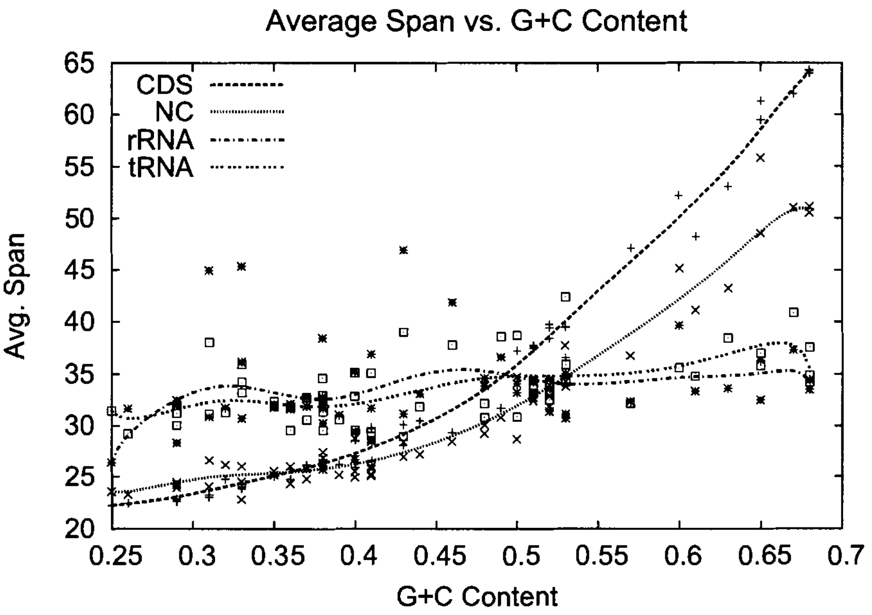


Figure 5.4: Average Stem-loop *span* vs. G+C content. The corresponding data is located in Appendix A.5

and/or the internal loops. Recall that the parameters which guide the search for the stem-loops restrict the size of the hairpin loops, bulges, and internal loops (Table 3.2). Therefore, the inherent nature of the stem-loop search largely precludes significant differences from emerging between the *bps* and *span* metrics.

Like the *bps* metric, there do not appear to be significant differences between the average *span* value in structural RNAs and those measured in their genomic counterparts to indicate this metric may act as a useful distinguishing feature. The *span* metric appears more stable in the rRNA and tRNA domains relative to their genomic counterparts (Table 5.4 and Figure 5.4). This is presumably attributable to the relatively stable G+C content levels in rRNA and tRNA.

5.3.1 Probability Distribution - *span*

The probability distribution for the *span* metric in *Borrelia burgdorferi*, NC_001318, is shown in Figure 5.5. The pattern displayed by these probability distributions explains the

Genomic Domain	**Average** *span*	**Standard Deviation**
CDS	33.94	12.34
NC	31.05	8.55
rRNA	33.54	4.02
tRNA	33.48	3.15

Table 5.4: Summary statistics for the average *span* for each genomic domain across the G+C content spectrum. The corresponding data is in Appendix A.5.

NC_001318 *span* **metric**	CDS	22.84 ± 7.58
	NC	23.97 ± 9.72
	rRNA	28.32 ± 11.79
	tRNA	29.82 ± 13.37

Table 5.5: Mean values for *span* metric in *Borrelia burgdorferi*, NC_001318, which has a global G+C content of 29%.

relatively high standard deviation listed in Table 5.5.

An analysis of the results obtained for the *span* metric in randomized or shuffled genomic sequences is presented in Appendix B.

5.4 Stem-loop Center-point Spacing - *cSpacing*

The *cSpacing* metric gauges the distance between stem-loops by measuring the average distance in nucleotides from the center-point of a given stem-loop to the center-points of its nearest non-overlapping upstream and downstream neighbours (Section 4.1.3). The average *cSpacing* values found in the genomic domains have been plotted against the *global* G+C content for each of the genomes in our working set. Figure 5.6 reveals some encouraging trends which are a remarkable improvement over the previous metrics.

In A+T rich genomes, the difference between rRNAs and its counterparts - CDS and NC - is quite significant (Figure 5.6). This discrepancy diminishes in more G+C rich genomes (Table 5.6). Yet, in all the previously described metrics, the difference between the average rRNA metric and the average of its counterparts decreases to zero with increasing

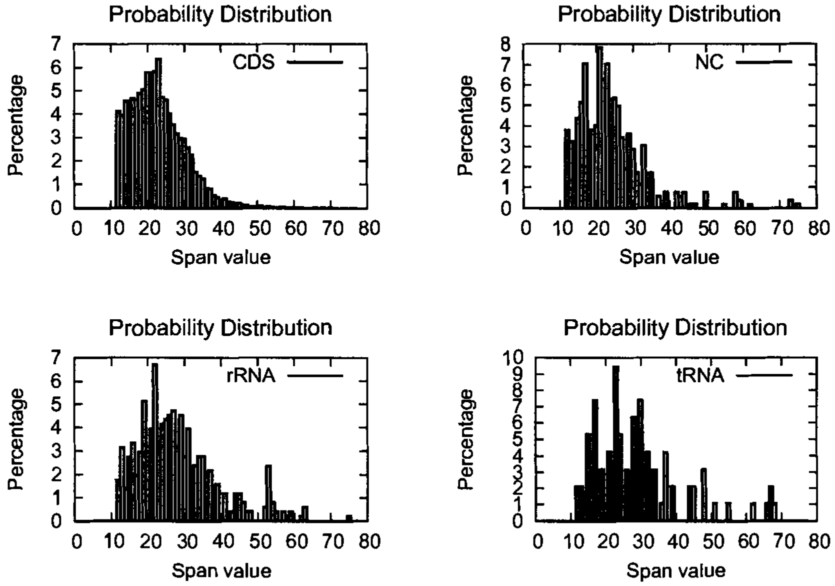


Figure 5.5: Probability distribution of the *span* metric in *B. burgdorferi*, NC_001318.

Accession Number	**G+C content**	**CDS** *cSpacing*	**NC** *cSpacing*	**rRNA** *cSpacing*
NC_003366	0.29	107.86	149.72	44.98
NC_003869	0.38	66.02	72.25	43.48
NC_000916	0.50	46.36	55.80	43.71
NC_000919	0.53	46.50	47.50	43.29
NC_002927	0.68	55.50	50.74	42.63

Table 5.6: The results presented depicted the diminishing discrepancy in the *cSpacing* metric as more G+C rich genomes are studied. This table is a short excerpt of the results presented in Appendix A.6

global G+C content - i.e. the rRNA plot intersects with those of its genomic counterparts (Figures 5.1, 5.2, and 5.4). In contrast, the rRNA *cSpacing* metric does *not* intersect with its counterparts - NC and CDS (Figure 5.6). Removing the tRNA values from Figure 5.6 helps to make this more clear - see Figure 5.7.

Another noteworthy feature regarding Figure 5.7 relates to variance. Relative to the CDS and NC regions, there is less variance in the average rRNA *cSpacing* metric across the *entire* G+C content spectrum (Table 5.7). The stability in *cSpacing* seems more striking than observed in the *bps* or *span* metrics. The low degree of variance is presumably related to conserved G+C content levels. More discussion on variability is found below in the description of the *cSpacing* probability distribution and later in Section 5.9.

Earlier, it was postulated that stem-loops might occur with higher frequency (i.e. lower average *cSpacing*) in regions which code for structural RNAs compared to their genomic counterparts. The results depicted in Figure 5.7 suggest that this may be true for rRNAs. However, when the G+C content in rRNAs and their counterparts is essentially equivalent (i.e. 50-55%), the difference in the *cSpacing* values between these groups is negligible (Table 5.6 and Appendix A.6).

5.4.1 Probability Distribution - *cSpacing*

Above, the *cSpacing* metric values measured in 58 genomic sequences was found to be significantly more stable in the rRNA domains in comparison to their counterparts - including tRNAs. This suggests that in just one of these genomes the standard deviation in the rRNA *cSpacing* metric could also be lower than that observed in its genomic counterparts. This

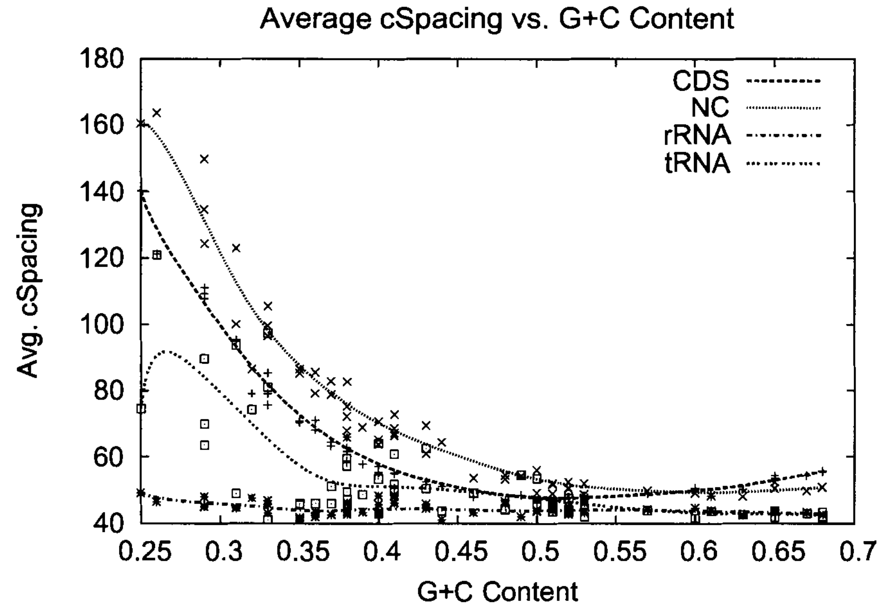


Figure 5.6: Stem-loop *cSpacing* vs. G+C content. The corresponding data is presented in Appendix A.6.

Genomic Domain	Average *cSpacing*	Standard Deviation
CDS	62.81	21.05
NC	72.72	28.86
rRNA	44.48	1.90
tRNA	53.43	16.00

Table 5.7: Summary statistics for the average *cSpacing* for each genomic domain across the G+C content spectrum. The corresponding data is presented in Appendix A.6.

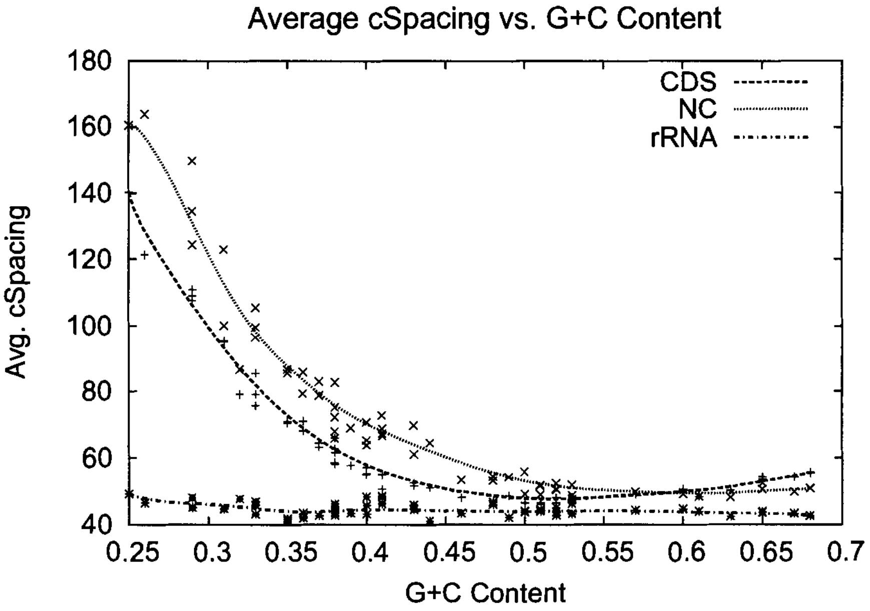


Figure 5.7: Stem-loop *cSpacing* vs. G+C content - tRNA values omitted. The corresponding data is presented in Appendix A.6.

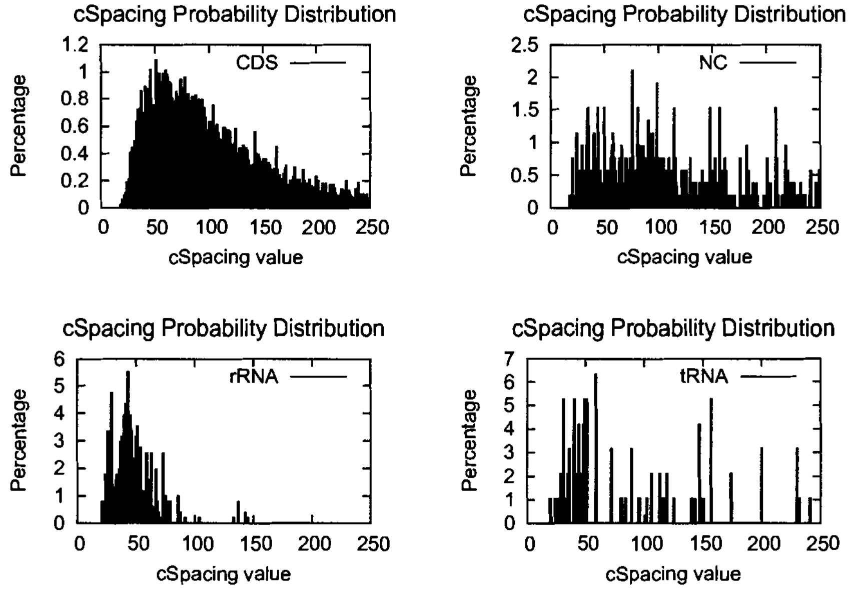


Figure 5.8: Probability distribution of the *cSpacing* metric in *B. burgdorferi*, NC_001318.

is indeed seen in the probability distribution and statistics on *B. burgdorferi*, NC_001318, shown in Figure 5.8 and Table 5.8.

An analysis of the results obtained for the *cSpacing* metric in shuffled genomic sequences is presented in Appendix B.

5.5 Stem-loop Foot Spacing - *fSpacing*

The foot spacing metric (*fSpacing*) is measured as the average distance to the foot of the nearest non-overlapping upstream and downstream stem-loops (Section 4.1.4). Trends evident in the *fSpacing* metric are similar to those described for the *cSpacing* metric. There is a remarkably low degree of *fSpacing* variance in the rRNA genes compared to their counterparts. There are some important differences between these metrics, however. Unlike *cSpacing*, the *fSpacing* rRNA plot intersects the CDS plot at roughly 50-55% G+C content (Figure 5.9). This rRNA plot also intersects the NC plot at roughly 60% G+C

NC_001318 ***cSpacing*** **metric**	CDS	111.04 ± 72.92
	NC	134.73 ± 107.06
	rRNA	47.94 ± 20.63
	tRNA	89.45 ± 63.12

Table 5.8: Mean values for *cSpacing* metric in *B. burgdorferi*, NC_001318, which has a global G+C content of 29%.

Genomic Domain	**Average** *fSpacing*	**Standard Deviation**
CDS	34.84	25.30
NC	46.33	31.98
rRNA	16.39	3.16
tRNA	25.95	16.45

Table 5.9: Summary statistics for the average *fSpacing* for each genomic domain across the G+C content spectrum. The corresponding data is located in Appendix A.7.

content. A statistical summary is presented in Table 5.9.

These results suggest the *cSpacing* metric has an advantage over the *fSpacing* metric since the rRNA *cSpacing* values do *not* intersect with their genomic counterparts (Figures 5.6 and 5.9). Why does this difference exist between two similar metrics? Recall, the *cSpacing* metric includes size or length attributes while the *fSpacing* metric excludes them. The *bps* and *span* metrics - both of which relate to the stem-loop size - revealed that the average values in the rRNAs domain display a tendency to diverge from the CDS and NC values at high G+C content levels (Figures 5.2 and 5.4). Therefore, by excluding the length attribute, the average *fSpacing* metric values in the CDS and NC domains lose their propensity to diverge from the average rRNA *fSpacing* value at these high G+C content levels. This likely explains why the differences between rRNA and its counterparts decreases to zero in the *fSpacing* metric but not in the *cSpacing* metric.

5.5.1 Probability Distribution - *fSpacing*

The probability distribution and statistics for *B. burgdorferi*, NC_001318, are depicted in Figure 5.10 and Table 5.10. The variance in the rRNA genes is considerably lower than its

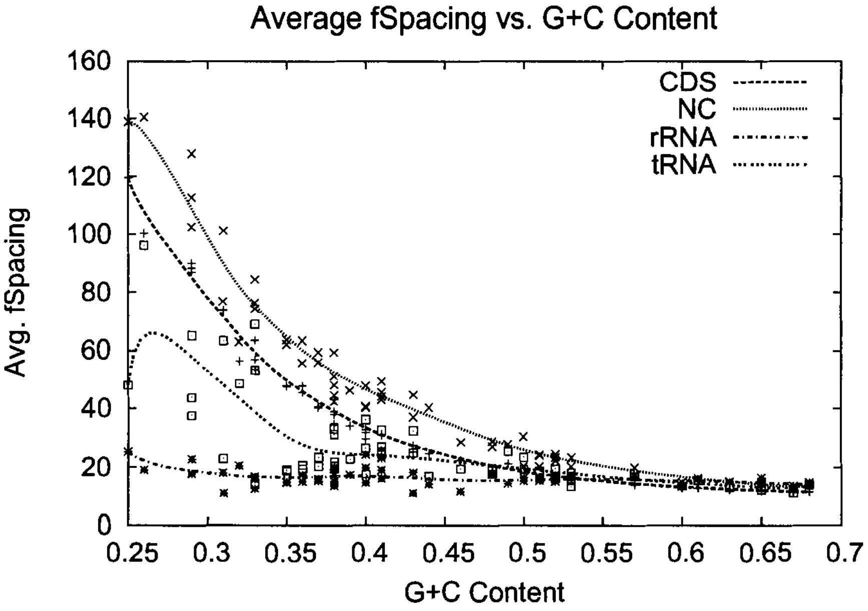


Figure 5.9: Average *fSpacing* vs. G+C content. The corresponding data is presented in Appendix A.7.

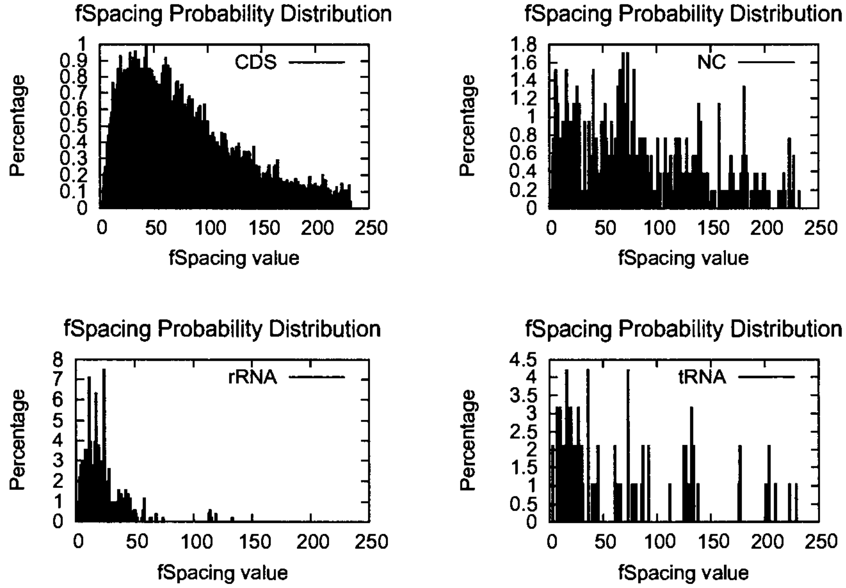


Figure 5.10: Probability distribution of the *fSpacing* metric in *B. burgdorferi*, NC_001318.

counterparts - including tRNAs. This may help to explain why the degree of variance seen in rRNA *fSpacing* across all 58 bacterial genomes is remarkably lower than the levels of variance seen in their genomic counterparts (Figure 5.9).

An analysis of the results obtained for the *fSpacing* metric in shuffled genomic sequences is presented in Appendix B.

5.6 Why rRNAs Outperform tRNAs

In these experiments, our goal is to evaluate the effectiveness of several stem-loop metrics in identifying structural RNAs along a genomic sequence. Interestingly, the *cSpacing* and *fSpacing* metrics perform significantly better on rRNAs than they do on tRNAs. Instinctively, one is prompted to ask why? The stem-loop search parameters are the likely reason. They were devised through cursory study of rRNA secondary structures. It seems the stem-loops which form in tRNAs are less likely to meet these search constraints. The results lend

NC_001318 *fSpacing* **metric**	CDS	89.88 ± 73.08
	NC	112.74 ± 107.37
	rRNA	22.74 ± 19.90
	tRNA	65.07 ± 62.29

Table 5.10: Mean values for *fSpacing* metric in *B. burgdorferi*, NC_001318, which has a global G+C content of 29%.

support of this line of reasoning. Consider that tRNAs are typically 70 to 90 nucleotides long [1]. Yet, the probability distribution for the *fSpacing* and *cSpacing* metrics in tRNAs depict stem-loops which are more than 200 nucleotides apart (Figures 5.8 and 5.10). These results may seem perplexing until one considers that tRNAs are commonly positioned side-by-side in a genomic sequence. The presence of tandem tRNAs explains why stem-loops in tRNA regions are sometimes divided by over 200 nucleotides. This suggests that the search parameters are not tailored to find the stem-loops which comprise tRNA structures.

The next section describes how metrics were combined in hopes of finding a *set* of metrics where the average rRNA metric is significantly divergent from its genomic counterparts across the entire G+C content spectrum.

5.7 Combined Metric - (*cSpacing* × *bps*)

The first combined metric takes the *cSpacing* and *bps* values for each stem-loop and multiplies them: (*cSpacing* × *bps*). The results are shown in Figure 5.11. They might be fairly described as unsteady or mixed. On the negative side, the rRNA plot intersects multiple times with the CDS and NC plots between 35-50% G+C content. This likely results from the proximity of their mean values and the high standard deviations associated with them (Table 5.11 and Appendix A.8). This is also supported by the probability distribution shown in Figure 5.12. On a positive note, the average (*cSpacing* × *bps*) values in rRNA appear markedly discrepant from their counterparts at approximately 50-55% *global* G+C. Such a strong discrepancy at this G+C content level has not been observed in any of the other metrics nor in the base composition method described previously. Another noteworthy feature regarding Figure 5.11 is the strong tendency for the average (*cSpacing* × *bps*) values

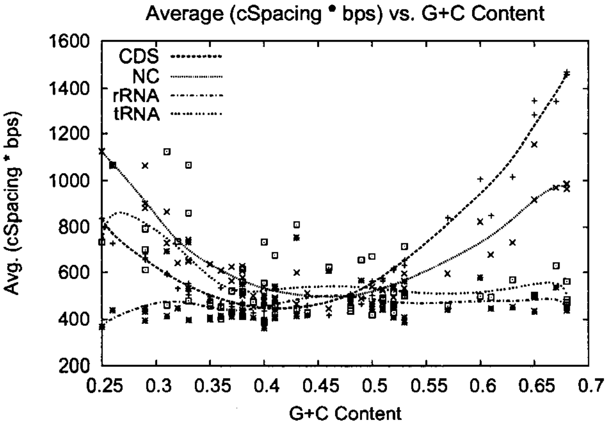


Figure 5.11: Avg. $\left(cSpacing \times bps\right)$ vs. G+C content. The corresponding data is presented in Appendix A.8.

to diverge from rRNA values in G+C rich genomes. Earlier, this same trend was noted in the *bps* and *span* metrics (Figures 5.2 and 5.4). Although less profound, this trend was also seen between the rRNA *cSpacing* metric and its counterparts (Figure 5.7). In retrospect, it appears that combining these metrics amplified several trends - both positive and negative - which were identified in their individual constituents.

Genomic Domain	**Average** $\left(cSpacing \times bps\right)$	**Standard Deviation**
CDS	591.81	231.68
NC	612.14	176.68
rRNA	437.05	70.17
tRNA	537.42	158.74

Table 5.11: Summary statistics for the average $\left(cSpacing \times bps\right)$ for each genomic domain across the G+C content spectrum. The corresponding data is presented in Appendix A.8.

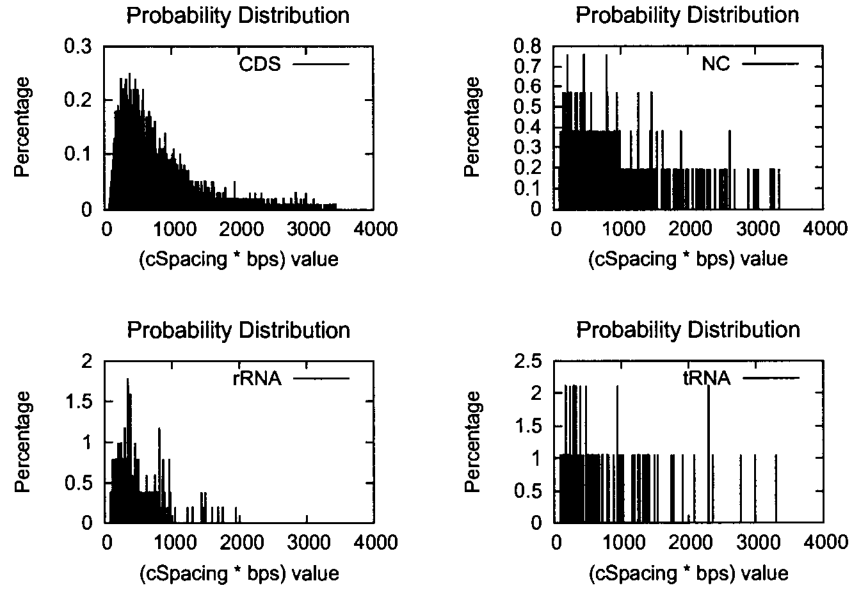


Figure 5.12: Probability distribution of the $\left(cSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318.

5.7.1 Probability Distribution - $\left(cSpacing \times bps\right)$

The probability distributions for the $\left(cSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318, cover a wide range of values (Figure 5.12). They illustrate the high degree of variance in this metric. The statistics on *B. burgdorferi* are shown in Table 5.12. As with previous metrics, the degree of variance in rRNA is less than what is observed in its counterparts.

An analysis of the results obtained for the $\left(cSpacing \times bps\right)$ metric in shuffled genomic sequences is presented in Appendix B.

NC_001318 $\left(cSpacing \times bps\right)$ **metric**	CDS	684.36 ± 525.65
	NC	1066.61 ± 782.05
	rRNA	438.89 ± 318.80
	tRNA	1069.02 ± 1077.31

Table 5.12: Mean values for $\left(cSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318 which has a global G+C content of 29%.

Genomic Domain	**Average** $\left(fSpacing \times bps\right)$	**Standard Deviation**
CDS	266.63	121.85
NC	349.57	208.05
rRNA	150.45	22.15
tRNA	258.82	221.23

Table 5.13: Summary statistics for the average $\left(fSpacing \times bps\right)$ for each genomic domain across the G+C content spectrum. The corresponding data is located in Appendix A.9.

5.8 Combined Metric - $(fSpacing \times bps)$

The final metric that was tested combines the $fSpacing$ and the bps metrics by multiplying them - $\left(fSpacing \times bps\right)$. The average for each of the respective genomic domains is plotted in Figure 5.13. The results are encouraging. The lines of fit for respective genomic domains suggest that the rRNA plot does not intersect its counterparts - CDS and NC. However, there is one instance where the average $\left(fSpacing \times bps\right)$ in rRNA regions is greater than the CDS and NC values. This lone culprit, NC_002935, has a G+C content of 53% (Appendix A.9). Nonetheless, these results suggest that this combined metric may be more apt at identifying rRNAs than either of their constituents acting alone - especially when the *global* G+C content approaches 50–55% (Figures 5.2 and 5.9). In Chapter 6, the limited benefits attained by using the $\left(fSpacing \times bps\right)$ metric over the $fSpacing$ metric are further documented. In addition, the degree of variance observed in the rRNA $\left(fSpacing \times bps\right)$ values over the 58 genomes is remarkably lower than those of its genomic counterparts (Figure 5.13 and Table 5.13).

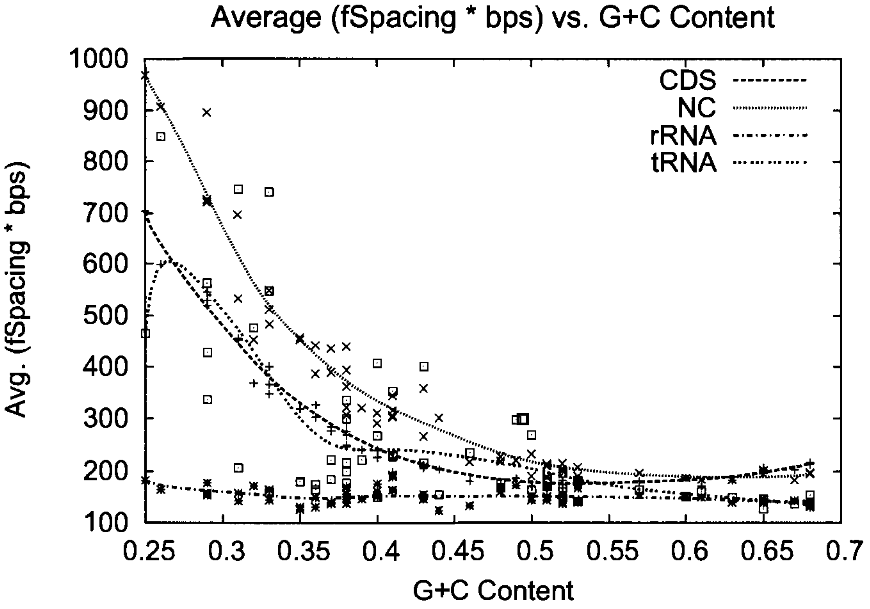


Figure 5.13: Average $\left(fSpacing \times bps\right)$ vs. G+C content. The corresponding data is presented in Appendix A.9.

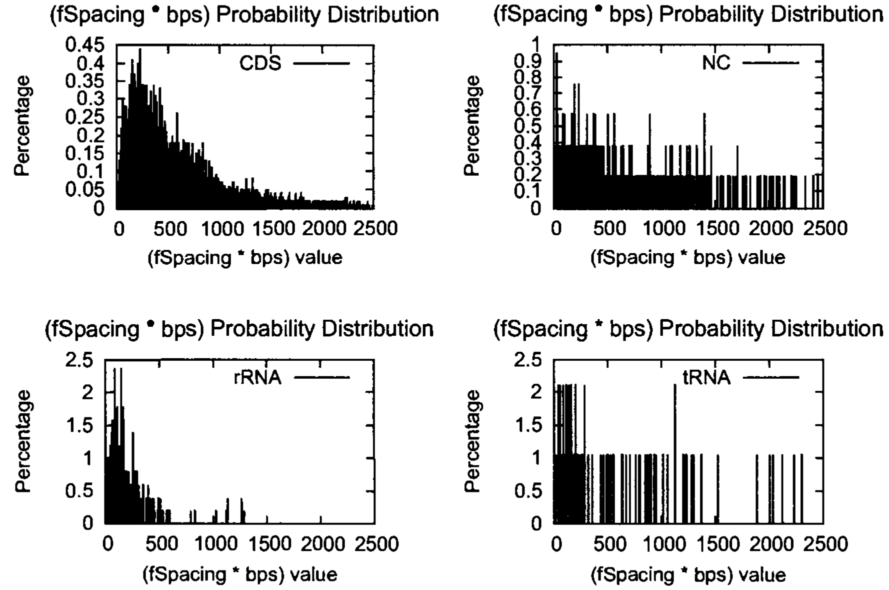


Figure 5.14: Probability distribution of the $\left(fSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318.

5.8.1 Probability Distribution - $\left(fSpacing \times bps\right)$

The probability distribution and statistics for the $\left(fSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318, are shown in Figure 5.14 and Table 5.14, respectively. The results above suggest that multiplying the $fSpacing$ and bps metrics helps to amplify differences in rRNA and its counterparts. The caveat, however, is that there is a high degree of variance in the $\left(fSpacing \times bps\right)$ values for each genomic domains in a given sequence (Table 5.13). Interestingly, the mean $\left(fSpacing \times bps\right)$ values in rRNA over the 58 sequences in our test set display a relatively low degree of variance.

An analysis of the results obtained for the $\left(fSpacing \times bps\right)$ metric in shuffled genomic sequences is presented in Appendix B.

NC_001318 $\left(fSpacing \times bps\right)$ **metric**	CDS	546.68 ± 493.65
	NC	725.77 ± 769.67
	rRNA	176.38 ± 182.06
	tRNA	551.11 ± 583.26

Table 5.14: Mean values for $\left(fSpacing \times bps\right)$ metric in *B. burgdorferi*, NC_001318 which has a global G+C content of 29%.

5.9 Variance in Stem-loop Spacing

One of the intriguing observations reported here is the remarkable stability in the *cSpacing* and *fSpacing* metrics in rRNA genes over the genomes in our test set which, incidentally, have a wide range of G+C content levels. These observations support the notion that structural RNA genes conserve structural information. Remarkably, this seems to be manifested in our admittedly simple stem-loop metrics.

It is interesting to speculate on the reasons behind the stability observed in the spacing metrics especially given the rRNA genes are located in 58 different bacterial genomes. One explanation is that the stability in the spacing metrics merely results from the stability in the G+C content in the rRNA genes. However, why do these rRNA genes tend to favour a specific base composition? This is likely related to the structural integrity provided by GC base pairs. Importantly, research by *Wang et al.* indicates that regions which are unpaired in the final RNA structure also tend to conserve base composition [64].

It is possible that rRNA transcripts favor an equilibrium of sorts. Such an equilibrium, it is postulated, might foster an environment conducive to attaining the final RNA structure. Presumably, an overabundance of stem-loops would increase the fraction of those which have to be dismantled to allow certain segments to "properly" pair-up for the final structure. This situation would seemingly impede an RNA transcript's natural propensity to form the intended final structure. Likewise, a resistance to forming stable stem-loops might prevent distant segments of the transcript from coming into close proximity. This, in turn, could create an environment unbecoming of the intended or final structure. Finding a balance between being too resistant to folding and being overtly folded may improve the ability of an rRNA transcript to efficiently reach its intended final structure. The consistency observed in rRNA G+C content and in rRNA spacing metrics lend support to this theory.

5.10 Chapter Review

Up to this point, several different stem-loop metrics have been proposed and examined. Initially, our suspicions were that stem-loops would occur on average with higher frequency and tend to be longer in structural RNAs when compared to their genomic counterparts. Our suggestions, however, have only proven to be partially correct. The results attained using the *fSpacing* metric suggest that stem-loops occur at a higher frequency in rRNA domains, generally, only when the structural RNA G+C content is greater than the *global* G+C content. Similarly, the average *bps* and *span* metric values of stem-loops in structural RNAs are longer, generally, only when the structural RNA G+C content is greater than the *global* G+C content. Hence, over the entire *global* G+C content spectrum, the *bps*, *span*, and *fSpacing* metrics each present a juncture where the values observed in structural RNAs and in their counterparts are equal. The *cSpacing* metric improves upon this in that the average rRNA *cSpacing* values are less than those of their genomic counterparts across the entire G+C content spectrum. However, when the G+C content level in rRNAs and its counter parts is roughly equal, there is little disparity in the *cSpacing* values for structural RNAs and their genomic counterparts. Various metrics were combined in an effort to amplify the disparity between structural RNAs and their genomic counterparts. This resulted in a limited degree of success.

Interestingly, the average stem-loop metric values recorded in structural RNA regions was marked by a relatively low degree of variance in genomes across the entire G+C content spectrum. This stability is presumably related to the resilient G+C content levels which are characteristic of structural RNAs. It is postulated that RNA transcripts may foster an equilibrium which is favourable to the intended RNA structure and to folding efficiently.

Chapter 6 reports the observations made when stem-loop metrics are used to locate regions which code for rRNA genes in various genomes without the help of a sequence map.

Chapter 6

Results - Locating Ribosomal RNA genes

This chapter extends the findings described in Chapter 5 by more closely examining the ability of chosen stem-loop metrics to locate rRNAs along a genomic sequence. The task in the experiments which follow is to identify rRNA genes along a genomic sequence by using stem-loop metrics and applying a statistical hypothesis test. A brief description of statistical inference and hypothesis testing is located in Section 4.2.2 on page 52.

There are a few key reasons for choosing to pursue rRNA genes along the sequence rather than tRNA genes. These reasons relate to the observations recently presented in Chapter 5 and to the inherent nature of the hypothesis test. Figures 5.6, 5.9, and 5.13 suggest that the discrepancy between the average stem-loop metrics in rRNA regions and their counterparts (i.e. CDS and NC) is greater than the discrepancy observed between tRNAs and their counterparts. The search parameters were tailored with rRNA genes in mind; this seems to explain why stem-loop metrics work less well on tRNAs (Section 5.6). Also, it is common practice to apply the CLT to sample sizes, N, of 30 or more [12]. tRNAs typically have 3–4 stem-loops in their final structure. Hence, samples which include no less than 30 adjacent stem-loops along the sequence would presumably overshoot the boundaries of a tRNA genes unless numerous tRNA genes occur in tandem. For these reasons, the experiments which follow examine the ability to uncover rRNAs rather than tRNAs.

6.1 Experimental Design

To find rRNA genes, the experiments which follow apply the average rRNA metric values reported in Tables 5.7, 5.11, and 5.13. Essentially, samples of adjacent stem-loops are taken along the entire length of the sequence. The mean or average metric value in these samples is compared to the rRNA values in the aforementioned tables using the hypothesis test. If the sample mean falls within the 95% confidence interval, the stem-loop at the center of the sample is classified as "structural RNA".

The first step involves finding the stem-loops along the input sequence. The average metric value is calculated by sampling the area around each stem-loop along the sequence. For instance, suppose the sample size is set to 51 stem-loops. The sample encompasses 25 stem-loops upstream plus 25 stem-loops downstream to the current stem-loop ($25+25+1 = 51$). Then, this sample mean is tested against the average *rRNA* metric value from one of the aforementioned tables. When numerous adjacent stem-loops fall *inside* the confidence interval this is manifested as a block of hits or "structural RNA" on the graphs which follow.

In later experiments, the sample sizes have been modified to improve the performance of the statistical test. Granted that increases to sample size can improve the results under some conditions, such modifications to the statistical test come with pitfalls. A larger sample spans a larger region of the sequence. The risk is that the sampled region may be larger than the targeted structural RNA genes. As a result, smaller structural RNAs may be overlooked and the periphery of large structural RNAs may not be included in the "hit". These concepts will become more clear as the results which follow are presented.

6.2 Using the *cSpacing* Metric to Find rRNA Genes

What follows are several examples where the *cSpacing* metric is used to find rRNA genes in the bacterial genomes. Note that the *global* G+C content level increases with each successive genome. Consequently, the task in finding the rRNA genes becomes progressively more difficult.

In the figures which follow, the aim is to correctly classify rRNAs as "structural RNA". Graphically, we are looking for the "hits" (labeled "S" for candidate structural RNA) to coincide with the positions where the rRNA genes (labeled "R"). What we are looking to avoid is a rampant number of "hits" where rRNA genes are not located. As sequences with a *global* G+C content approaching 50% are tested, the limited effectiveness of the stem-loop *cSpacing* metric emerges. This is manifested by an increasing number of false positives. False positives are regarded as a preponderance of hits or "structural RNAs" in the results graphs. The assumption is that structural RNAs are not likely to make-up more than 10-15% of any given genome.

Methanococcus maripaludis S2, NC_005791, has a *global* G+C content of 33%. The average G+C content in the rRNA genes is 54%. In Figure 6.1, there are two rRNA genes which reside in the genomic segment which is presented. Recall, junctures where the average *cSpacing* metric falls *inside* the 95% confidence interval are classified as "structural RNA" (see the caption in Figure 6.1). The bottom graph in Figure 6.1 is noteworthy since it depicts a sharp rise in *local* G+C content where the rRNAs reside. The correspondence between the hits and the rRNAs indicates that their locations have been correctly identified.

Compared to *M. maripaludis*, *Bacillus cereus* ATCC 10987, NC_003909, has a slightly higher *global* G+C content which measures 36%; the rRNA genes have an average G+C content of 52%. Figure 6.2 shows that the statistical test successfully delineates the rRNA genes. This is attributable to the relatively large difference in the average stem-loop spacing between the rRNA domains and its counterparts and to the sharp difference in *local* G+C content between the rRNA genes and its genomic counterparts.

Chlamydia muridarum, NC_002620, has a *global* G+C content of 40%; the average G+C content in rRNA genes is 49%. The results of its analysis are depicted in Figure 6.3. Although some false positives appear, the rRNAs are reasonably well delineated from their counterparts.

Figure 6.4 depicts the results for *Escherichia coli*, NC_004431, which has a *global* G+C content of 50%. The average G+C content in the rRNA genes is 53%. Similarly, Figure 6.5 depicts the results for *Corynebacterium diphtheriae*, NC_002935. It has a *global* G+C content of 53% and an average G+C content in the rRNA genes is 54%. The abundance of false positives seen in these genomes is attributable to the proximity of the mean rRNA *cSpacing* metric and the mean values of its counterparts (Table 5.6 and Figure 5.7). One can view this problematic scenario as opposing probability distributions which overlap with one another (Figure 6.6). Note that the sample size was increased in these genomes to narrow the probability distributions of the sample means. However, there is too much overlap between the opposing domains to make this work effectively.

Recall from Figure 5.7 that as the *global* G+C content level surpasses roughly 54% the

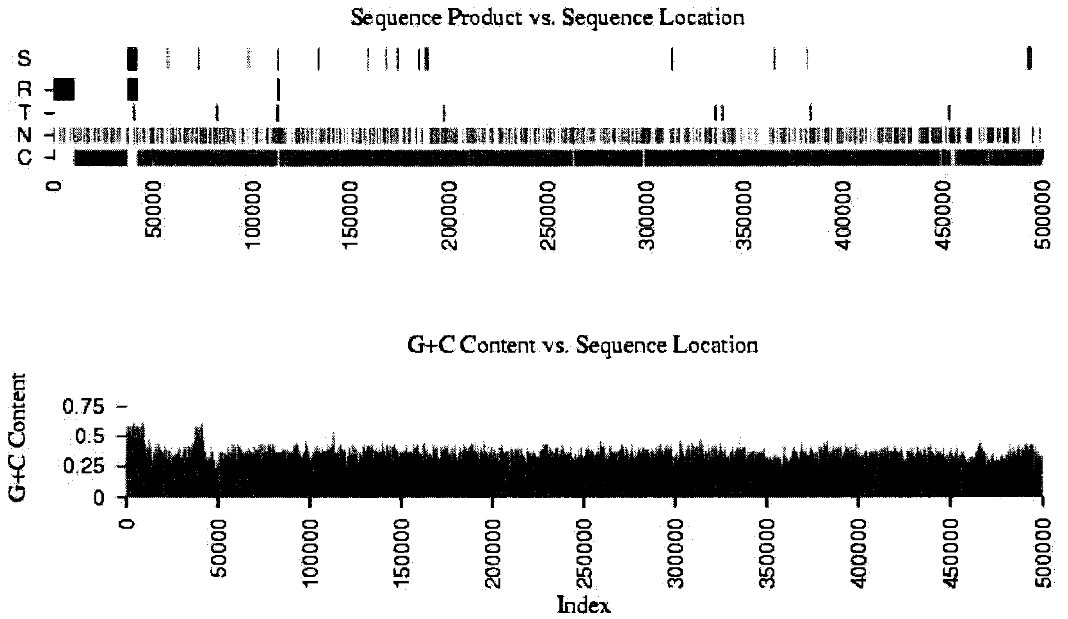


Figure 6.1: NC_005791: *Methanococcus maripaludis S2* [0 .. 500,000], Sample size 51 stems; *Global* G+C = 33%; Threshold setting = 95%; Metric: *cSpacing*; This graph depicts the ability of the *cSpacing* metric to delineate rRNA genes in the bacterial genome. The top graph depicts a map of the sequence which divides it into 4 broad categories. They include coding sequences (C), noncoding DNA (N), tRNA (T), and rRNA (R). In addition, the segments classified as candidate structural RNA by the statistical test are labeled "S". An ideal result is marked by a strong correlation between the candidate structural RNAs - i.e. the "hits" - and the location of the rRNA genes. The bottom graphs displays variations in G+C content along the sequence. In the interest of space, the sequences are graphed in 500,000 nucleotide segments.

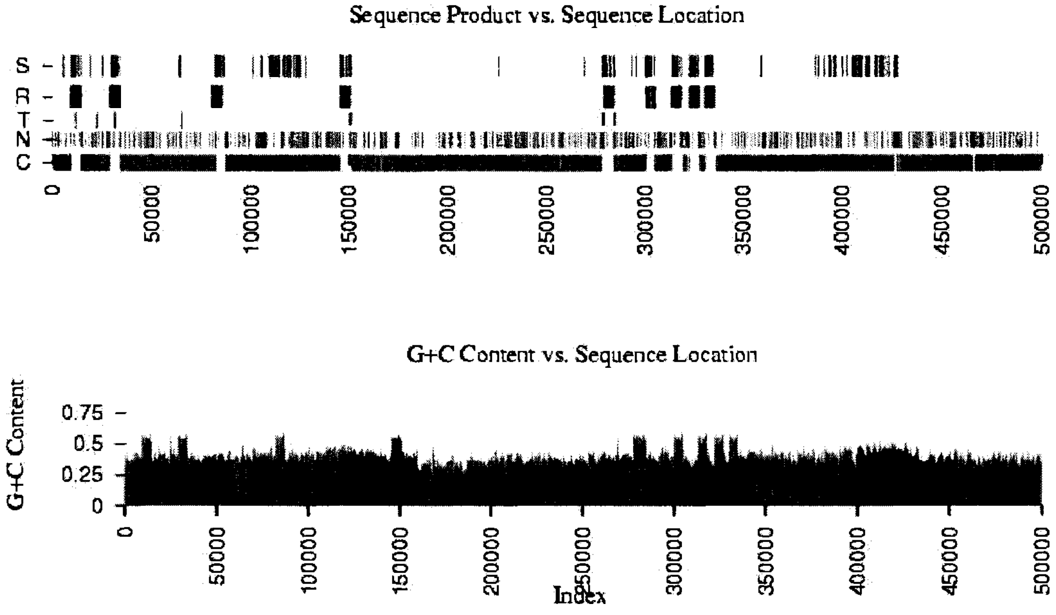


Figure 6.2: NC_003909 (*Bacillus cereus* ATCC 10987) [0 .. 500,000], Sample size 51 stems; *Global* G+C content= 36%; Threshold setting = 95%; Metric: (*cSpacing*); This graph depicts the ability of the *cSpacing* metric to delineate rRNA genes in the bacterial genome. The top graph depicts a map of the sequence which divides it into 4 broad categories. They include coding sequences (C), noncoding DNA (N), tRNA (T), and rRNA (R). In addition, the segments classified as candidate structural RNA by the statistical test are labeled "S". An ideal result is marked by a strong correlation between the candidate structural RNAs - i.e. the "hits" - and the location of the rRNA genes. The bottom graphs displays variations in G+C content along the sequence. In the interest of space, the sequences are graphed in 500,000 nucleotide segments.

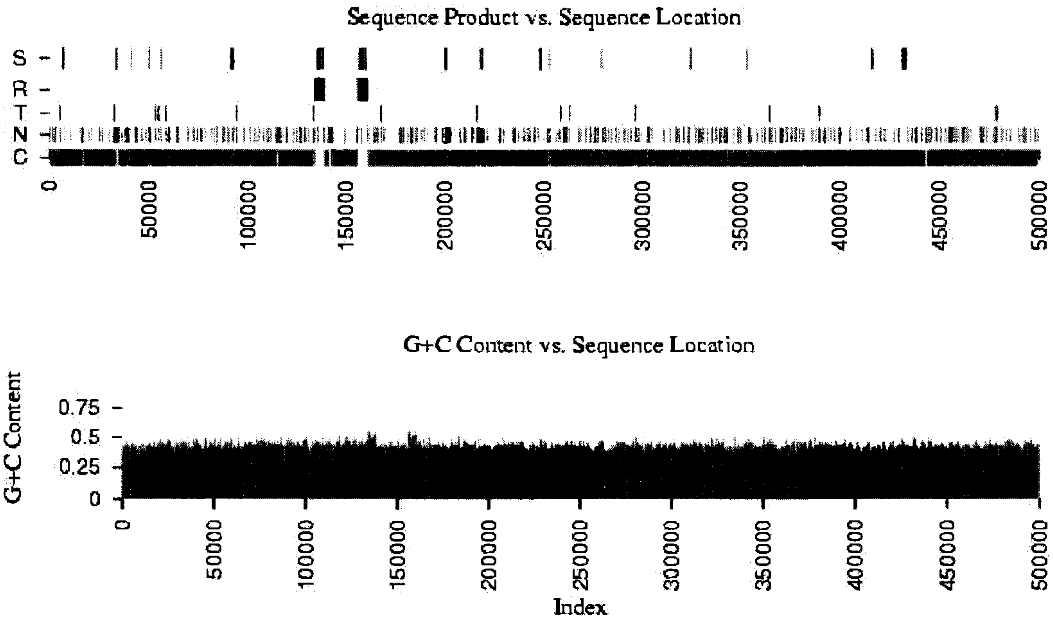


Figure 6.3: NC_002620, *Chlamydia muridarum*, [0 .. 500,000], Sample size 121 stems; *global* G+C = 40%; Threshold setting = 95%; Metric: (*cSpacing*). A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

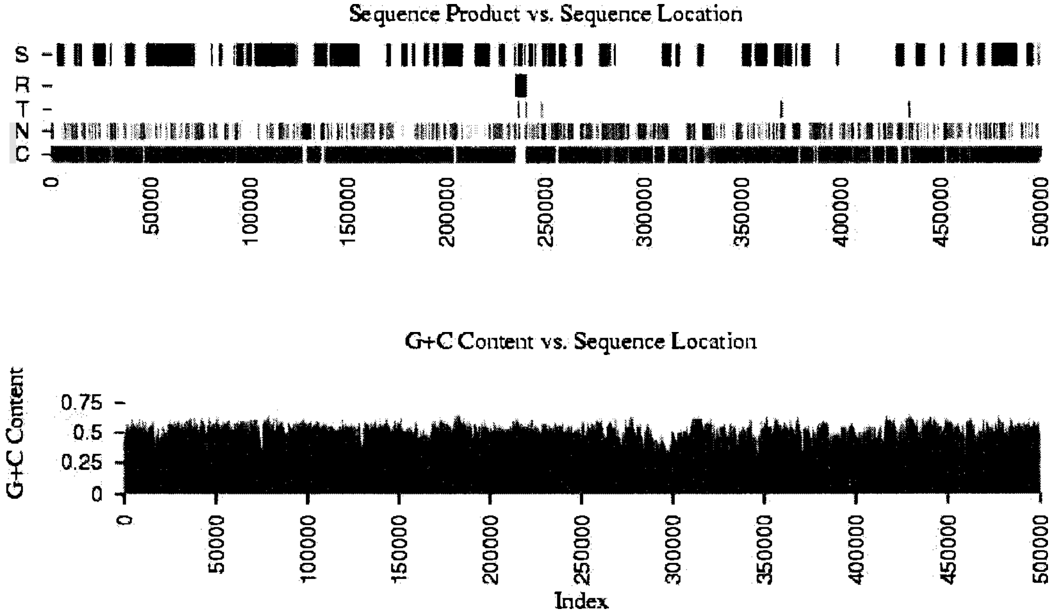


Figure 6.4: NC_004431, *Escherichia coli* CFT073, [0 .. 500,000] , Sample size 501 stems; *global* G+C = 50%; Threshold setting = 95%; Metric: (*cSpacing*). A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

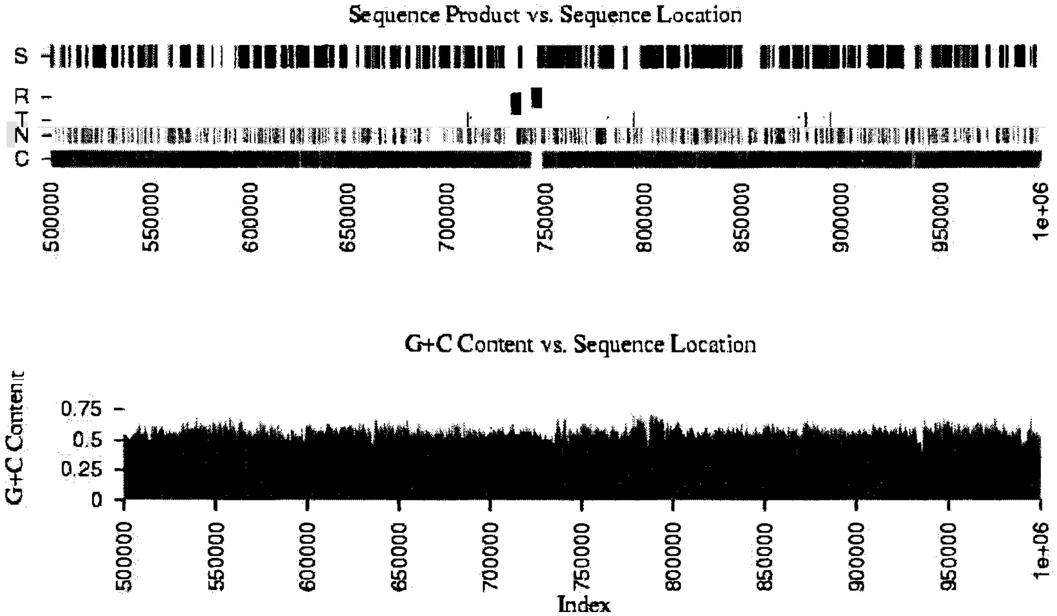


Figure 6.5: NC_002935, *Corynebacterium diphtheriae*, [500,000 .. 1,000,000], Sample size 301 stems; *global* G+C = 53%; Threshold setting = 95%; Metric: (*cSpacing*). A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

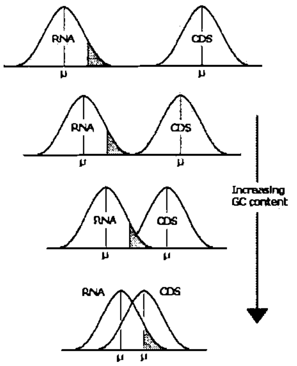


Figure 6.6: Overlapping Probability Distributions of Normally Distributed sample means. As the G+C content approaches 0.50 the population means of the *cSpacing* metric move into close proximity. This results in overlapping probability distributions as depicted above. Consequently, a given threshold in one probability distribution may no longer exclude sample means which have arisen from an overlapping counterpart. More false positives emerge as a result.

mean rRNA *cSpacing* value and that of its genomic counterparts begin to diverge. This is manifested in the decreasing number of false positives in genomes where the *global* G+C content is greater than 54%. *Chlorobium tepidum*, NC_002932, has a *global* G+C content of 57%; its rRNA genes have an average G+C content of 52%. Notably, it has fewer false positives than NC_002935 (Figure 6.7). Similarly, *Bordetella bronchiseptica*, NC_002927, has a *global* G+C content of 68% compared to an average 54% G+C content level in its rRNA genes. Its results are depicted in Figure 6.8.

6.3 Using Combined Stem-loop Metrics to Find rRNA Genes

The results presented suggest the largest hurdle to the stem-loop *cSpacing* metric occurs when the *global* G+C content approaches the average G+C content found in rRNA genes - roughly 50-54%. In Sections 5.7 and 5.8, stem-loop metrics where combined in hopes of distancing the average rRNA metric value from the average metric values found in their genomic counterparts. As mentioned earlier in Section 5.7, the plots resulting from these combinations have both positive and negative aspects (Figures 5.11 and 5.13). The following examples shed more light on the ability of these "combined" metrics to identify rRNA genes

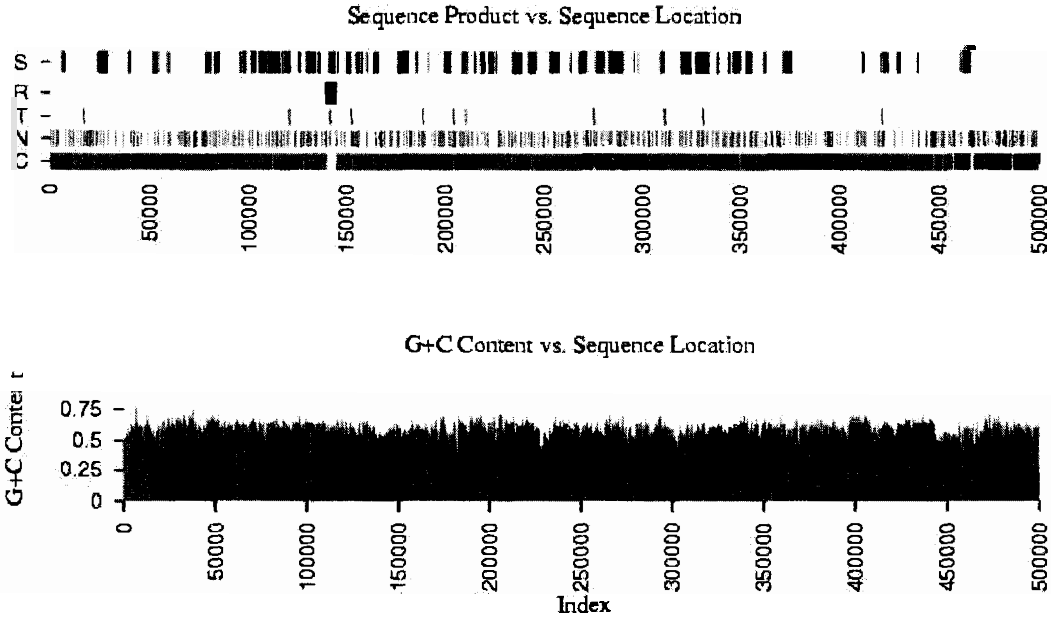


Figure 6.7: NC_002932, *Chlorobium tepidum* TLS, [0 .. 500,000], Sample size 501 stems; *global* G+C = 57%; Threshold setting = 95%; Metric: (*cSpacing*). A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

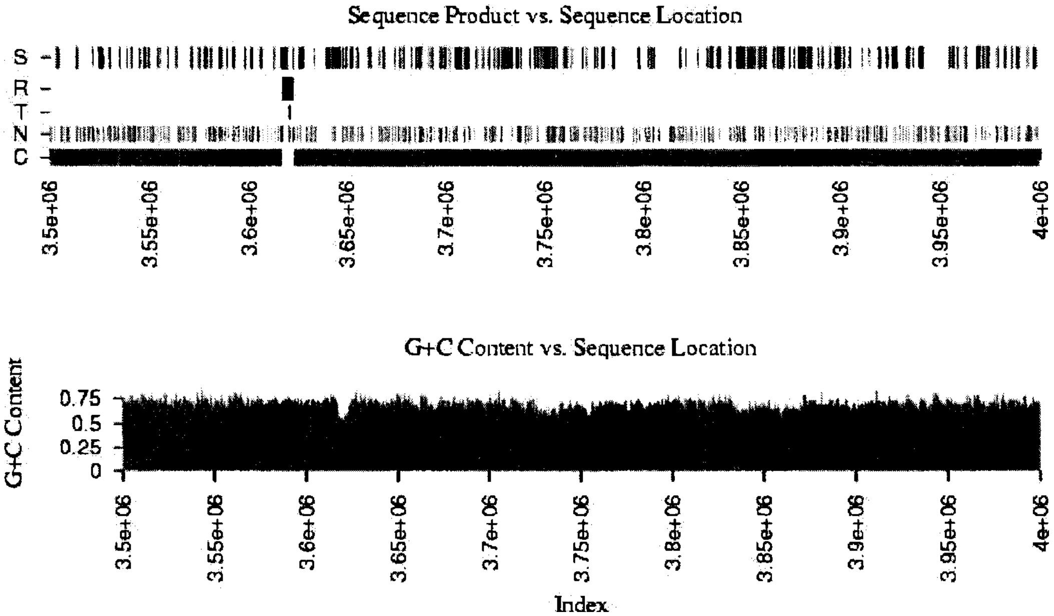


Figure 6.8: NC_002927, *Bordetella bronchiseptica* RB50, [3,500,000 .. 4,000,000], Sample size 151 stems; *global* G+C = 68%; Threshold setting = 95%; Metric: (*cSpacing*). A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

when there is a negligible difference between the *global* G+C content and the G+C content in rRNA genes.

In *Treponema pallidum*, NC_000919, both the *global* G+C content and rRNA G+C content is 53%. Using the $\left(fSpacing \times bps\right)$ metric, the regions which code for rRNA genes are identified, however, there are too many false positives (Figure 6.9).

In *Corynebacterium diphtheriae*, NC_002935, the *global* G+C content is 53% and the G+C content in the rRNA genes is 54%. The results for the $\left(cSpacing \times bps\right)$ metric are shown in Figure 6.10. Although there are several false positives, their frequency has notably decreased in comparison to the *cSpacing* metric analysis on the same genome (Figure 6.5).

In *Salmonella typhimurium* LT2, NC_003197, the *global* G+C content is 52%. The rRNA genes have an average G+C content of 54%. The $\left(cSpacing \times bps\right)$ identifies the rRNA genes yet too many false positives are present (Figure 6.11).

In *Chlorobium tepidum* TLS, NC_002932, the *global* G+C content is 57%; the G+C content in the rRNA genes is 52%. The $\left(cSpacing \times bps\right)$ metric identifies the rRNA genes in the segment depicted (Figure 6.12). The results are encouraging, however, several false positives persist. It appears that the difference in G+C content levels has helped to reduce the false positive frequency. A similar situation emerges in *Bifidobacterium longum*, NC_002939. The *global* G+C content is 61%; the average G+C content in the rRNA genes is 55% (Figure 6.13). The $\left(cSpacing \times bps\right)$ metric locates the rRNA genes with fewer false positives in comparison to previous examples.

The results clearly document that negligible differences in *global* G+C content and rRNA G+C content tend to obscure rRNA genes. This is problematic for our stem-loop metric approach and for the base composition method. However, the results presented for the stem-loop method are quite encouraging - especially considering it is in its infancy. When the difference between the *global* G+C content and the rRNA G+C content is merely 1-2%, the $\left(cSpacing \times bps\right)$ metric correctly - in large part - eliminates roughly half of the sequence as being rRNA material (Figures 6.10 and 6.11). It seems unlikely that strictly adhering to a base composition method could accomplish this feat. A direct comparison would be helpful in lending support to this suggestion.

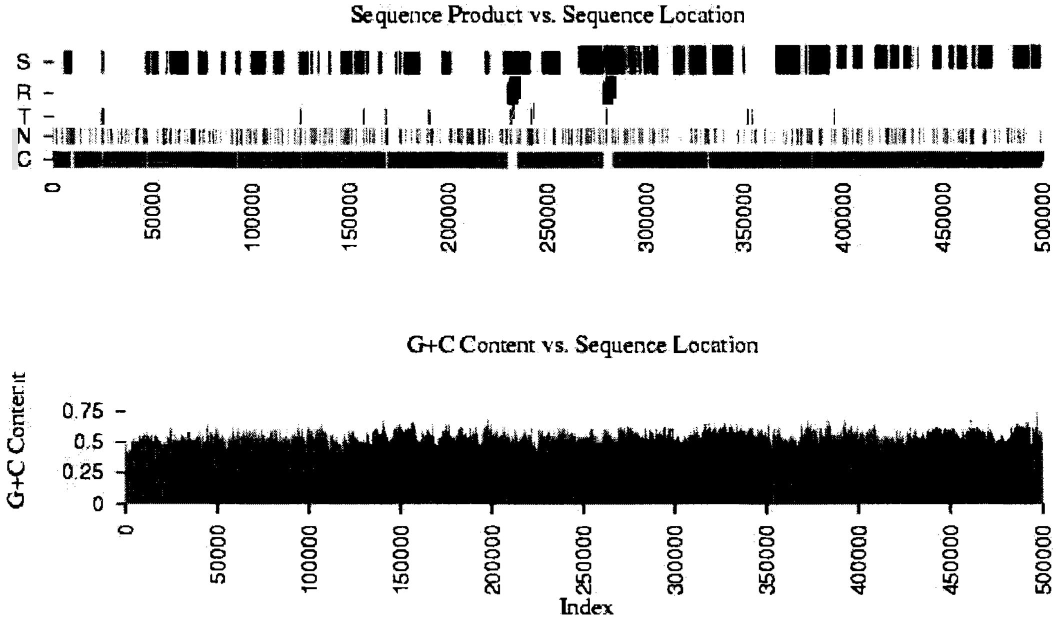


Figure 6.9: NC_000919, *Treponema pallidum*, [0 .. 500,000], Sample size 501 stems; *global* G+C = 53%; Threshold setting = 95% (2–sided test). Metric: $\left(fSpacing \times bps\right)$. A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

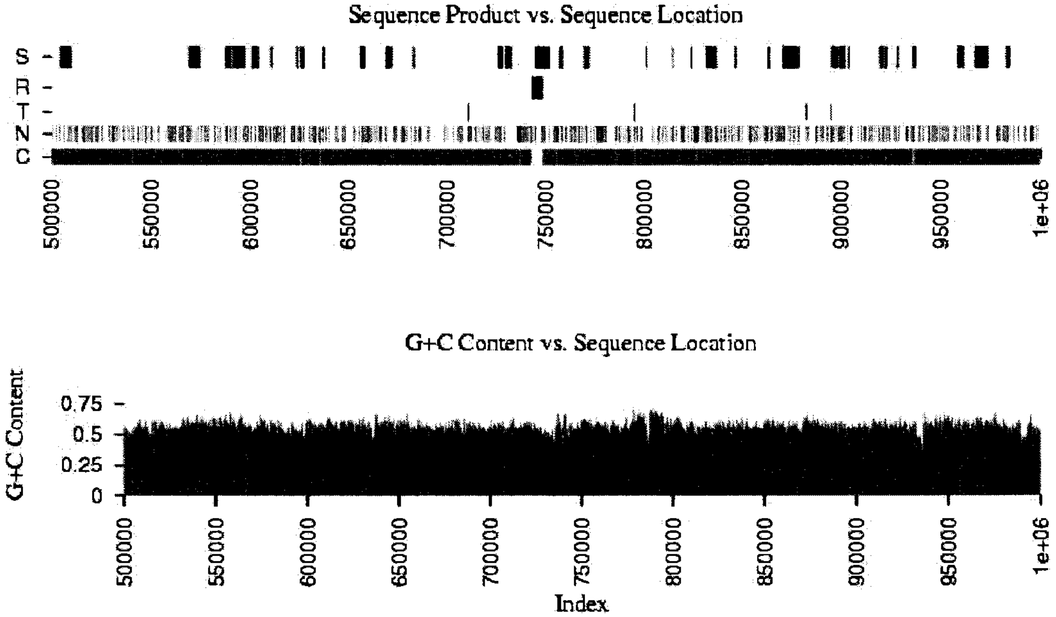


Figure 6.10: NC_002935, *Corynebacterium diphtheriae*, [500,000 .. 1,000,000], Sample size 601 stems; *global* G+C = 53%; Threshold setting = 95% (2–sided test). Metric: $\left(cSpacing \times bps\right)$. A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

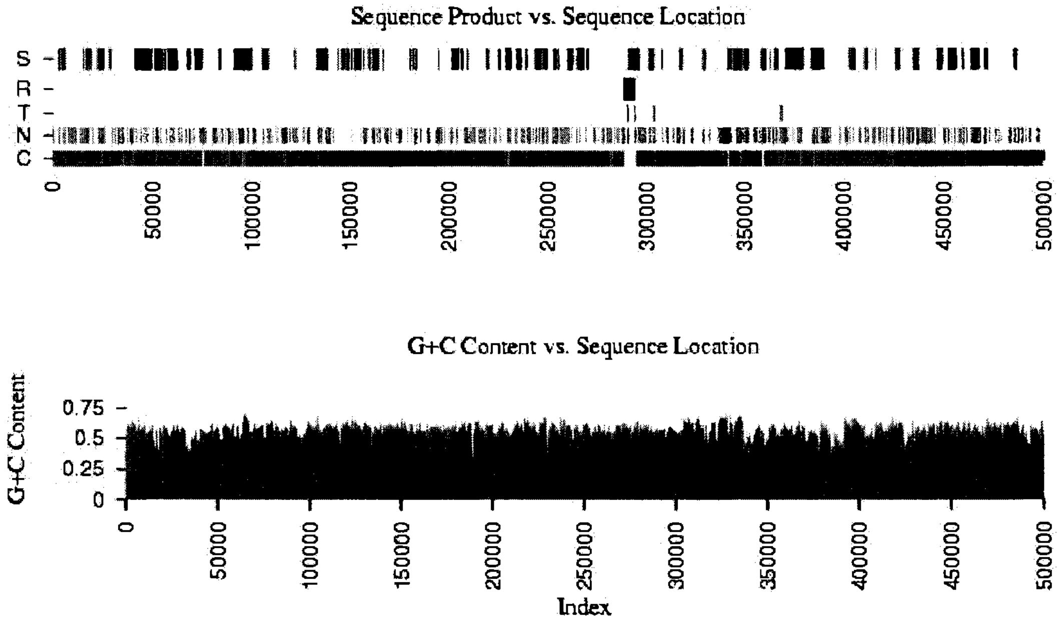


Figure 6.11: NC_003197, *Salmonella typhimurium* LT2 , [0 .. 500,000], Sample size 301 stems; *global* G+C = 52%; Threshold setting = 95% (2-sided test); Metric: $\left(cSpacing \times bps\right)$. A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

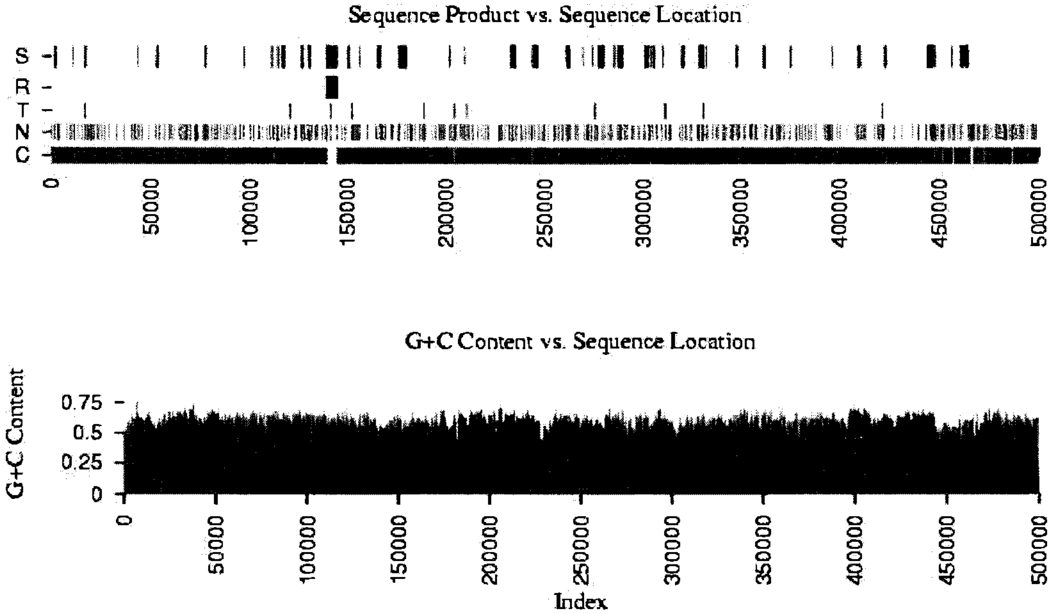


Figure 6.12: NC_002932, *Chlorobium tepidum* TLS, [0 .. 500,000], Sample size 251 stems; *global* G+C = 57%; Threshold setting = 95%; Metric: $\left(cSpacing \times bps\right)$. A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

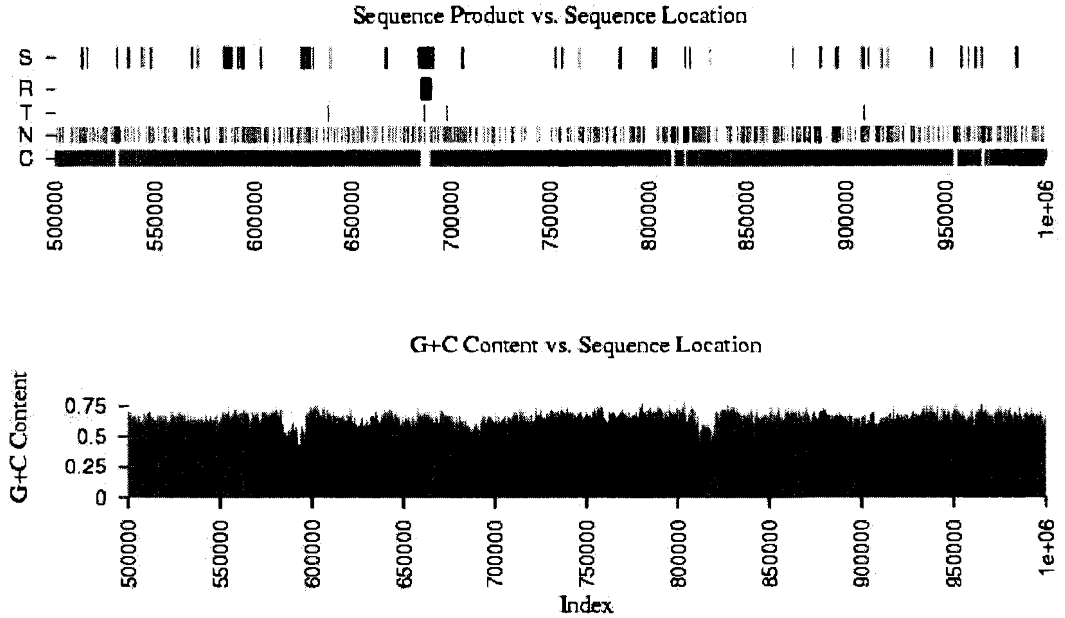


Figure 6.13: NC_002939, *Geobacter sulfurreducens* PCA, [500,000 .. 1,000,000], Sample size 201 stems; *global* G+C = 61%; Threshold setting = 95% (2-sided test); Metric: $\left(cSpacing \times bps\right)$. A detailed description of the information conveyed in these graphs can be found in Figure 6.2.

6.4 Chapter Review

Several stem-loop metrics were tested on 58 different bacterial genomes. The search parameters were tailored to rRNA secondary structures. As a result, the metrics were more studied for their ability to identify rRNA genes rather than tRNA genes.

The results indicate that the most difficult hurdle in identifying rRNA genes occurs when there is a negligible difference between *global* G+C content and rRNA G+C content. Stem-loop metrics were combined in an effort to overcome this hurdle. This was met with limited success. In many cases, a high incidence of false positives could not be overcome. Nonetheless, in sequences where the difference in G+C content levels between rRNA genes and their counterparts is 1-2%, this approach eliminates roughly half of the genome as structural RNA material with a promising degree of accuracy. The results reported here are encouraging especially considering this stem-loop based approach is one which has not been studied before.

Chapter 7 presents a number of possible avenues to proceed with further research.

Chapter 7

Suggestions for Future Research

Our observations show that stem-loop metrics can be used to identify rRNA genes across a wide range of G+C content levels. They also point to some of the key hurdles which need to be overcome. Nonetheless, the results suggest further study is warranted. This chapter presents some possible directions for future research.

7.1 Improve Accuracy

The stem-loop finding algorithm could be improved to include all the stem-loops which meet the search parameters. The current implementation does not store all of the qualifying stem-loops. One culprit is the `Select1PathFrom2` which is described in Section 3.4 on page 31. Inherently, this function discards variants of stems which meet the parameters but are tossed aside in favour of larger runs of consecutive base pairs. It would be prudent to keep all stem-loops which meet the search parameters. This would presumably improve the ability to identify stem-loops which are observed in secondary structure diagrams.

7.2 Improve Efficiency

Refinements could be made to improve the efficiency of the search algorithm. This includes eliminating redundancy in exploring the search space where bulges and internal loops are identified. Currently, the algorithm looks for a stretch of base pairs starting from each of the cells in the table depicted in Figure 3.21. The algorithm moves through the table from left to right and top to bottom finding stretches of base pairs along the way. When the algorithm shifts down to a lower row, it diligently considers every possible cell to look for a stretch of base pairs. Suppose a path of adjacent base pairs has already been located starting from a higher row in the table. In this situation, the algorithm (working through a lower row of cells) may unnecessarily evaluate previously discovered base pairs along a given diagonal (or path). This redundancy can be addressed by applying a map to indicate which cells have already been checked. Alternatively, this search space could be examined by using dynamic programming.

7.3 Parameter Optimization

Earlier, the ramifications of using search parameters tailored to rRNA genes were discussed. It would be interesting to explore how varying the search parameters affects the results. For illustration, the minimum GC base pair constraint and the maximum GU base pair constraint were eliminated (Table 3.2). Hence, the minimum fraction of GC base pairs is ≥ 0.0 and the maximum fraction of GU base pairs is ≤ 1.0. The results for the *bps*, *cSpacing*, and *fSpacing* metrics under these new constraints are shown in Figures 7.1, 7.2, and 7.3 respectively. Generally, the average spacing in rRNA and tRNAs is less than their genomic counterparts. However, the discrepancy between rRNAs and their genomic counterparts appears less than seen earlier under the initial set of search parameters. Hence, it seems the averages are not sufficiently distinct to identify rRNA genes without being overrun by false positives. Nonetheless, experimentation with various search parameters could be fruitful.

7.4 Improve Portfolio of Stem-loop Metrics

More study is required to devise new and improved stem-loop metrics which better distinguish rRNAs from their genomic counterparts. As mentioned previously, a stem-loop metric which is characterized by a low degree of variance in rRNA genes is advantageous. These advantages would be further extended if a set of stem-loop metrics were stable through not only rRNA genes but also through other unrelated structural RNA genes. This could help to uncover more intriguing structural RNA genes such as *XIST* and *H19*.

Also, it would be interesting to study improvements garnered by adopting motif sequences found in tetra-loops into the search engine.

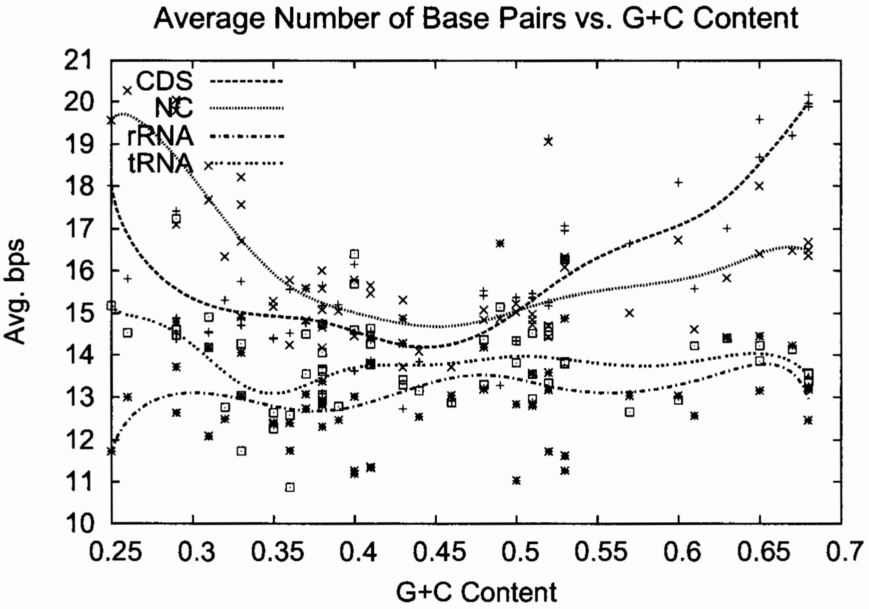


Figure 7.1: The average *bps* vs. G+C content when the minimum GC base pair content is set to 0.0% and the maximum GU base pair content is set to 1.0. The corresponding data is located in Appendix A.10.

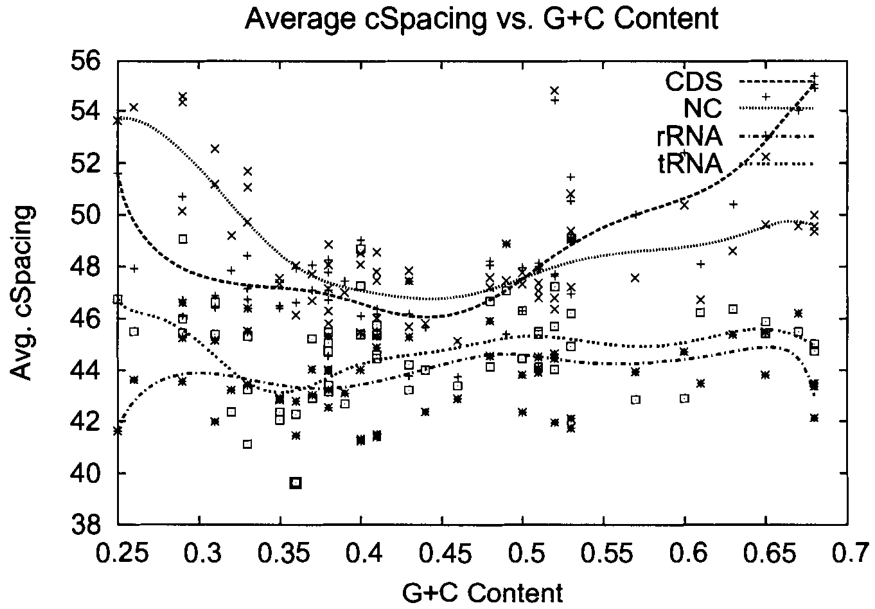


Figure 7.2: The average *cSpacing* vs. G+C content when the minimum GC base pair content is set to 0.0% and the maximum GU base pair content is set to 1.0. The corresponding data is located in Appendix A.11.

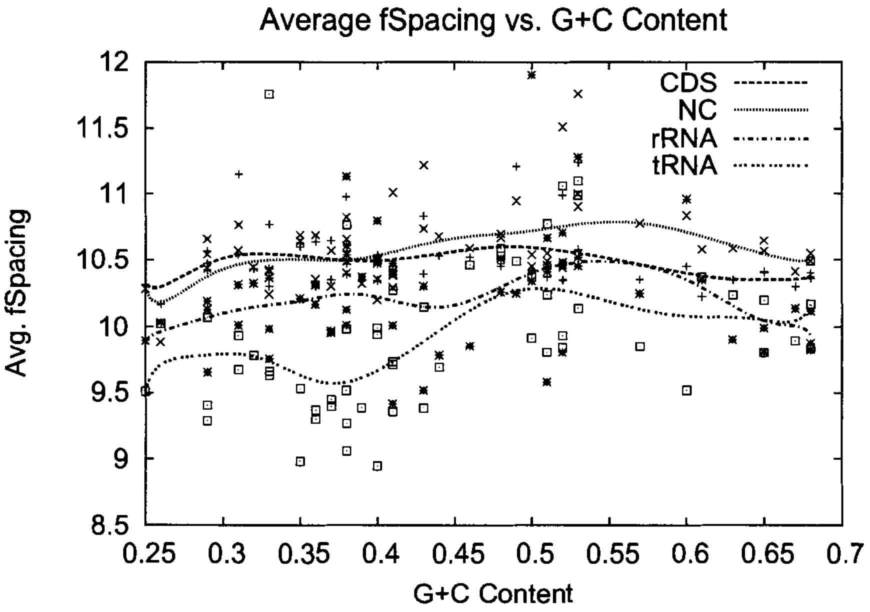


Figure 7.3: The average *fSpacing* vs. G+C content when the minimum GC base pair content is set to 0.0% and the maximum GU base pair content is set to 1.0. The corresponding data is located in Appendix A.12.

7.5 Analyze Both Strands

The current implementation of the search algorithm scans only 1 strand of the double-stranded DNA sequence. However, all the CDS, NC, rRNAs, and tRNAs are not necessarily located on this strand. Given this entire body of work is based on average valuations, it was assumed that the difference between the 2 strands would not be sufficient to invalidate our results. It is important to keep in mind that CDS and NC domains change from one sequence to the next. Likewise, rRNAs of various size of all pooled together to tabulate the average rRNA metric value. Hence, there is an inherent lack of precision when working with averages. Our interest is not in the absolute values for each of the respective genomic domains but more in the differences between them.

A more accurate and precise definition of the annotated genomes can be attained with the use of two 1-dimensional maps (i.e. 1 for each strand). It is anticipated the precision provided by such a map would not overcome the shortcomings of the stem-loop metrics. However, the results of such experiments would be interesting.

7.6 Collaborate With Wet Laboratory

Lastly, it would be interesting to pair-up with a wet lab to determine whether some of the "hits" uncovered by using stem-loops metrics are indeed novel structural RNA genes.

7.7 Chapter Review

A number of possible avenues for future efforts have been proposed. One of the key hurdles to overcome is compromised effectiveness of stem-loop metrics when the base composition in structural RNAs and their genomic counterparts is essentially uniform.

Chapter 8

Conclusion

This research project has introduced and examined a novel approach to locate regions which code for structural RNA genes along a genomic sequence. This approach is founded on the stem-loop - a recurrent substructure or component universally found in naturally occurring RNA secondary structures. The objective of this project was to study whether stem-loops could be useful in finding structural RNA genes. To meet this objective a multitude of programs and scripts had to be designed and implemented.

The primary algorithm searches a nucleotide sequence for stem-loops. Importantly, a set of search parameters was devised to identify stem-loops similar to those found in rRNA secondary structures. As the algorithm was being tested, 2 performance related issues emerged. First, the program, which was initially implemented in Python, took approximately 7 minutes to scan a 10^6 nucleotide sequence with a Intel Pentium $4^{©}$ 2.6 GHz processor. To overcome this limitation, the algorithm was implemented in C++. As a result, the time to scan 10^6 nucleotide had been reduced to about 6 seconds.

Further testing revealed that the C++ implementation would virtually grind to a halt on sequences longer than roughly 8×10^6 nucleotides. The problem, it turned out, was memory related; the program was consuming all 1.4 GB of RAM memory and the 500 MB of disk swap too. This problem was rectified by dynamically monitoring the memory consumption and disposing of data which was no longer needed as the algorithm continued along the sequence. As a result, the C++ implementation typically uses less that 300 MB of RAM and the CPU runs at virtually full capacity though the duration of the scan. At this juncture, a program required to scan numerous genomic sequences was in place.

To study the efficacy of stem-loops in pursuing structural RNA genes, a number of stem-loop metrics were devised. With the help of annotated sequences downloaded from the NCBI public databases, the average values for these metrics could be calculated over the various genomic domains along the sequences. These domains include protein-coding sequences (CDS), non-coding DNA (NC), rRNAs, and tRNAs. The goal was to determine whether the average metric values in these genomic regions were significantly discrepant. Significant differences in these averages would suggest regions of interest could be delineated along a genomic sequence. In scanning each of the annotated genomes, the average stem-loop metric for each of the respective genomic domains was recorded.

A total of 58 bacterial genomes were arbitrarily selected. They represented a diversity of *global* G+C content levels ranging from 25% to 68%. After scanning each of the genomes, the data was analyzed for trends related to changes in *global* G+C content. This was done with the use of graphs. In doing so, each metric had its own graph. The *global* G+C content was plotted along the x-axis, the metric value was plotted along the y-axis. The metric values recorded in the CDS, NC, rRNA, and tRNA domains for each genome are positioned vertically to one another since they are all derived from a single genome with a single *global* G+C content value. The vertical distance between the metric values for the various genomic domains provides an indication of how discrepant they are at a given *global* G+C content level. This allowed us to study how changes in *global* G+C content levels affect the metric values. Importantly, it also revealed how discrepant the structural RNAs are from their genomic counterparts in terms of stem-loop metrics. Our goal was to find a set of stem-loop metrics where the average value in the structural RNAs differ significantly from their genomic counterparts across the entire G+C content spectrum.

Several stem-loop metrics were evaluated. They included: *bps*, *span*, *cSpacing*, and *fSpacing*. The search parameters were devised to target stem-loops occurring in rRNA structures. This was manifested in the results. The stem-loop metrics were more apt at distinguishing rRNA genes than tRNA genes. Neither the *bps* nor the *span* metric appeared capable of distinguishing rRNAs from their genomic counterparts. The *cSpacing* and *fSpacing* metrics showed more promise. However, when *global* G+C content and rRNA G+C content is roughly equivalent, there is little disparity in their spacing metric values. This did not bode well for distinguishing structural RNAs from their genomic counterparts. In an effort to overcome this obstacle, combined metrics were devised. They include: $\left(fSpacing \times bps\right)$ and $\left(cSpacing \times bps\right)$. There was a slight improvement in the results. However, the discrepancy between the rRNA domain values and those of their counterparts was not consistently significant.

The final set of experiments more closely examined the ability of stem-loop metrics to identify structural RNAs occurring along a genomic sequence. The question at this juncture was: Given the average stem-loop metric value in the rRNA genes, can they be identified without referring to a sequence map? The *cSpacing* metric successfully identified rRNA genes in genomic sequences with less than roughly 42% *global* G+C content and more than roughly 62% *global* G+C content. The *fSpacing* metric performed well in genomes with less than roughly 42% *global* G+C content. The previous experiments foreshadowed this outcome. These metrics are not able to adequately delineate rRNA genes when their base composition and that of their genomic counterparts is essentially uniform. There was a small improvement when metrics were combined. In some sequences where G+C content is fairly uniform, the $\left(cSpacing \times bps\right)$ metric correctly cast aside roughly half of the genome. More study is required to find a set of metrics that performs well across the entire G+C content spectrum.

In several bacterial genomes which span several millions of nucleotides, this novel approach accurately identifies rRNAs. Hopefully, this stem-loop centered approach will mature to find not only rRNA genes but to find previously undiscovered structural RNAs as well.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* Garland Publishing, New York, 4^{th} edition, 2002.

[2] Hatim T. Allawi and Jr. John SantaLucia. Thermodynamics and nmr of internal gt mismatches in dna. *Biochemistry*, 36:10581–10594, June 1997.

[3] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. GenBank. *Nucleic Acids Research*, 31(1):23–27, 2003.

[4] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. GenBank: update. *Nucleic Acids Research*, 32:D23–D26, 2004. Database issue.

[5] Emily Bernstein, Amy A. Caudy, Scott M. Hammond, and Gregory J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 408(6818):295–296, Jan 2001.

[6] Camilynn I. Brannan, Elizabeth Claire Dees, Robert S. Ingram, and Shirley M. Tilghman. The product of the *H19* gene may function as an RNA. *Molecular and Cellular Biology*, 10(1):28-36, 1990.

[7] Carolyn J. Brown, Andrea Ballabio, James L. Rupert, Ronald G. Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F. Willard. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304):38–44, Jan 1991.

[8] Richard Carter, Inna Dubchak, and Stephen Holbrook. A computational approach to identify genes for structural RNAs in genomic sequences. *Nucleic Acids Research*, 29:3928–3938, 2001.

[9] Jih-Hsiang Chen, Shu-Yun Le, Bruce Shapiro, Kathleen Currey, and Jacob Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Computer Applications in the Biosciences*, 6:7–18, 1990.

[10] Jacob Z. Dalgaard and Roger A. Garrett. *The Biochemistry of Archaea (Archaebacteria)*, chapter Archaeal hyperthermophile genes, pages 535–563. Elsevier, 1993.

[11] Alain Deschênes and Kay C. Wiese. Using stacking-energies (INN and INN-HB) for improving the accuracy of RNA secondary structure prediction with an evolutionary algorithm - a comparison to known structures. In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*, volume 1, pages 598–606, Portland, Oregon, June 2004. IEEE Press.

[12] Jay L. Devore. *Probability And Statistics For Engineering and The Sciences*. Duxbury, California, USA, 5th edition, 2000.

[13] Sean R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2:919–929, Dec 2001. Review.

[14] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[15] Volkner A. Erdmann, Maciej Szymański, Abraham Hochberg, Nathan de Groot, and Jan Barciszewski. Collection of mRNA-like non-coding RNAs. *Nucleic Acids Research*, 27(1):192–195, 1999.

[16] James W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acid Research*, 10(17):5303–5318, 1982.

[17] Nicolas Galtier and J.R. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, 44:632–636, 1997.

[18] P.J. Grabowski, S.R. Seiler, and P.A. Sharp. A multicomponent complex is involved in the splicing of messenger RNA percursors. *Cell*, 42(1):345–353, Aug 1985.

[19] Alexander P. Gultyaev, F. H. D. van Batenburg, and Cornelis W. A. Pleij. The computer-simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250:37–51, 1995.

[20] Liyan He, Ryszard Kierzek, Jr. John SantaLucia, Amy E. Walter, and Douglas H. Runer. Nearest-neighbor parameters for gu mismatches: GU/UG is destabilizing in the contexts CGUG/GUGC, UGUA/AUGU but stabillizing in GGUC/CUGG. *Biochemistry*, 30:11124–11132, 1991.

[21] Stephen R. Holbrook and Sung-Hou Kim. RNA crystallography. *Biopolymers*, 44(1):3–21, 1997.

[22] L.M. Hsu, J. Zagorski, Z. Wang, and M.J. Fournier. Escherichia coli 6S RNA gene is part of a dual-function transcription unit. *Journal of Bacteriology*, 161(3):1162–1170, Mar 1985.

[23] L.D. Hurst and A.R. Merchant. High guanine-cytosine content is not an adaptation to high temperature in prokaryotes: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London - Biological Sciences*, 268:493–497, 2001.

[24] Alexander Hüttenhofer, Martin Kiefmann, Sebastian Meier-Ewert, John O'Brien, Hans Lehrach, Jean-Pierre Bachellerie, and Jürgen Brosius. Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO Journal*, 20:2943–2953, 2001.

[25] John Santa Lucia Jr., Ryszard Kierzek, and Douglas H. Turner. Stabilities of consecutive A.C, C.C, G.G, U.C, and U.U mismatches in RNA internal loops: Evidence for stable hydrogen-bonded U.U and C.C+ pairs. *Biochemistry*, 30:8242–8251, 1991.

[26] Samuel Karlin, Allan M. Campbell, and Jan Mrázek. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, 32:185–225, 1998. Review.

[27] Elzbieta Kierzek, Ewa Biala, and Ryszard Kierzek. Elements of thermodynamics in RNA evolution. *Acta Biochim. Pol.*, 48:485–493, 2001.

[28] James Kim, Amy E. Walter, and Douglas H. Turner. Thermodynamics of coaxially stacked helixes with GA and CC mismatches. *Biochemistry*, 35:13753–13761, 1996.

[29] Anders Krogh. Two methods for improving performance of a HMM and their application for gene finding. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 179–186, Menlo Park, CA, 1997. AAAI Press.

[30] J. Kurz, J. Lovely, S. Cubitt, and M.O. Krause. Distinct nuclear 7s RNAs hybridize to regulatory regions of two oncogenes. *Biochemistry Biophysics Research Communications*, 152(2):753–761, Apr 1988.

[31] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294:853–858, 2001.

[32] Eric C. Lai. MicroRNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics*, 30(4):363–364, Apr 2002.

[33] Nelson Lau, Lee Lim, Earl Weinstein, and David Bartel. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294:858–862, 2001.

[34] Shu-Yun Le, Jih-Hsiang Chen, Kathleen Currey, and Jacob Maizel. A program for predicting significant RNA secondary structures. *Computer Applications in the Biosciences*, 4:153–159, 1988.

[35] Rosalind Lee and Victor Ambros. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 294:862–864, 2001.

[36] Carl E. Longfellow, Ryszard Kierzek, and Douglas H. Turner. Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, 29:278–285, 1990.

[37] Arthur J. Lustig. Telomerase RNA: a flexible RNA scaffold for telomerase biosynthesis. *Current Biology*, 14(14):R565–567, Jul 2004.

[38] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.

[39] David H. Mathews and Douglas H. Turner. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, 41:869–880, 2002.

[40] May Meroueh and Christine S. Chow. Thermodynamics of RNA hairpins containing single internal mismatches. *Nucleic Acids Research*, 27(4), 1999.

[41] Brian R. Morton and Bernadette G. So. Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. *Journal of Molecular Evolution*, 50(2):184–193, February 2000.

[42] Adele Murrell, Sarah Heeson, and Wolf Reik. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genetics*, 36(8):889–893, August 2004.

[43] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*, 35:68–82, 1978.

[44] Adam E. Peritz, Ryszard Kierzek, Naoki Sugimoto, and Douglas H. Turner. Thermodynamic study of internal loops in oligoribonucleotides: Symmetric loops are more stable than asymmetric loops. *Biochemistry*, 30:6428–6436, 1991.

[45] Stephen Prata. *C++ Primer Plus*. Waite Group, U.S.A., 3^{rd} edition, 1998.

[46] Stanley Radel and Marjorie Navidi. *Chemistry*. West Publishing Company 1, St. Paul, MN., 1990.

[47] Ramaswamy Ramakrishna and Ramachandran Srinivasan. Gene identification in bacterial and organellar genomes using GeneScan. *Computers and Chemistry*, 23:165–174, 1999.

[48] Elena Rivas and Sean R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[49] Elena Rivas and Sean R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

[50] Daniel P. Romero and Elizabeth H. Blackburn. A conserved secondary stucture for telomerase RNA. *Cell*, 67(2):343–353, Oct 1991.

[51] Gary Ruvkun. Glimpses of a tiny RNA world. *Science*, 294:797–799, 2001. Review.

[52] John Santa-Lucia Jr. and Douglas H. Turner. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, 44:309–319, 1997.

[53] Peter Schattner. Searching for RNA genes using base-composition statistics. *Nucleic Acids Research*, 30(9):2076–2082, 2002.

[54] Susan J. Schroeder, Mark E. Burkard, and Douglas H. Turner. The energetics of small internal loops in RNA. *Biopolymers*, 52:157–167, 1999.

[55] Martin J. Serra and Douglas H. Turner. Predicting thermodynamic properties of RNA. *Methods in Enzymology*, 259:242–261, 1995.

[56] Frank Sleutels, Ronald Zwart, and Denise P. Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415(6873):810–813, Feb 2002.

[57] Dieter Söll, Susumu Nishiumura, and Peter B. Moore, editors. *RNA*. Elsevier Science Ltd., Kidlington, Oxford, 2001.

[58] Rodger Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.

[59] Gisela Storz. An expanding universe of noncoding RNAs. *Science*, 296:1260–1262, May 2002.

[60] Maciej Szymański, Volker A Erdmann, and Jan Barciszewski. Noncoding regulatory RNAs database. *Nucleic Acids Research*, 31(1):429–431, 2003.

[61] Ignacio Tinoco Jr. and Carlos Bustamante. How RNA folds. *Journal of Molecular Biology*, 293:271–281, 1999.

[62] F. H. D. van Batenburg, A. P. Gultyaev, and C. W. A. Pleij. An APL-programmed genetic algorithm for the prediction of rna secondary structure. *Journal of Theoretical Biology*, 174:269–280, 1995.

[63] Akiyoshi Wada and Akira Suyama. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Progress in Biophysics and Molecular Biology*, 47:113–157, 1986. Review.

[64] Huai-chun Wang and Donal A. Hickey. Evidence for strong selective constraint acting on the nucleotide composition of 16s ribosomal RNA genes. *Nucleic Acids Research*, 30(11):2501–2507, 2002.

[65] Zasha Weinberg and Walter L. Ruzzo. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, 20(Suppl. 1):i334–i341, 2004.

[66] Carl R. Woese, S. Winker, and Robin R. Gutell. Architecture of ribsomal RNA: Constraints on the sequence of "tetra-loops". *Proceeding fo the National Academy of Science*, 87:8467–8471, Nov. 1990.

[67] Michael Wolfinger. The energy landscape of RNA folding. Master's thesis, University of Vienna, 2001.

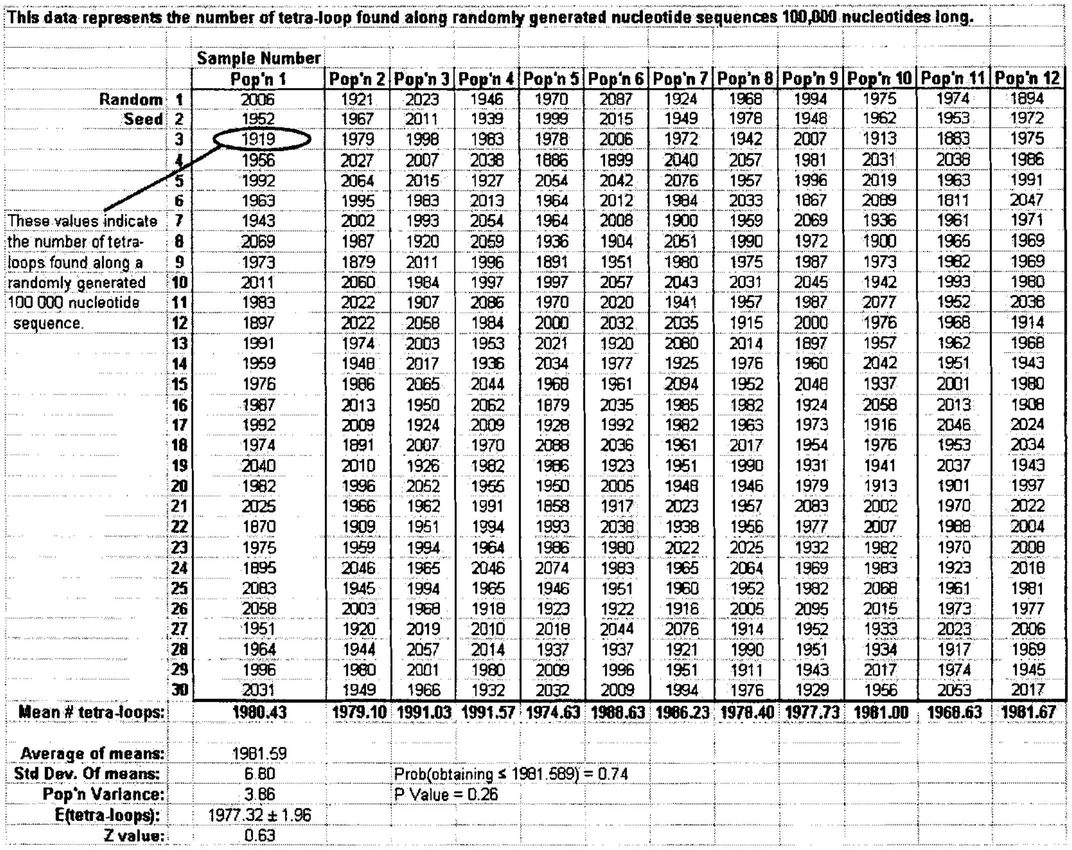
[68] Michael Zuker. Prediction of RNA secondary structure by energy minimization. In Annette M. Griffin and Hugh G. Griffin, editors, *Computer Analysis of Sequence Data*, pages 267–294. Humana Press Inc., July 1994.

[69] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[70] Micheal Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

Appendix A

Stem-loop Metric Data on Genomic Sequences



This data represents the number of tetra-loop found along randomly generated nucleotide sequences 100,000 nucleotides long.

		Sample Number											
		Pop'n 1	Pop'n 2	Pop'n 3	Pop'n 4	Pop'n 5	Pop'n 6	Pop'n 7	Pop'n 8	Pop'n 9	Pop'n 10	Pop'n 11	Pop'n 12
Random	1	2006	1921	2023	1946	1970	2087	1924	1968	1994	1975	1974	1894
Seed	2	1952	1967	2011	1939	1999	2015	1949	1978	1948	1962	1953	1972
	3	1919	1979	1998	1983	1978	2006	1972	1942	2007	1913	1883	1975
	4	1956	2027	2007	2038	1886	1899	2040	2057	1981	2031	2038	1986
	5	1992	2064	2015	1927	2054	2042	2076	1957	1996	2019	1963	1991
	6	1963	1995	1983	2013	1964	2012	1984	2033	1867	2089	1811	2047
These values indicate	7	1943	2002	1993	2054	1964	2008	1900	1959	2069	1936	1961	1971
the number of tetra-	8	2069	1987	1920	2059	1936	1904	2051	1990	1972	1900	1965	1969
loops found along a	9	1973	1879	2011	1996	1891	1951	1980	1975	1987	1973	1982	1969
randomly generated	10	2011	2060	1984	1997	1997	2057	2043	2031	2045	1942	1993	1980
100 000 nucleotide	11	1983	2022	1907	2086	1970	2020	1941	1957	1987	2077	1952	2038
sequence.	12	1897	2022	2058	1984	2000	2032	2035	1915	2000	1976	1968	1914
	13	1991	1974	2003	1953	2021	1920	2080	2014	1897	1957	1962	1968
	14	1959	1948	2017	1936	2034	1977	1925	1976	1960	2042	1951	1943
	15	1976	1986	2065	2044	1968	1961	2094	1952	2048	1937	2001	1980
	16	1987	2013	1950	2062	1879	2035	1985	1982	1924	2058	2013	1908
	17	1992	2009	1924	2009	1928	1992	1982	1963	1973	1916	2046	2024
	18	1974	1891	2007	1970	2088	2036	1961	2017	1954	1976	1953	2034
	19	2040	2010	1926	1982	1986	1923	1951	1990	1931	1941	2037	1943
	20	1982	1996	2052	1955	1950	2005	1948	1946	1979	1913	1901	1997
	21	2025	1966	1962	1991	1858	1917	2023	1957	2083	2002	1970	2022
	22	1870	1909	1951	1994	1993	2038	1938	1956	1977	2007	1988	2004
	23	1975	1959	1994	1964	1986	1980	2022	2025	1932	1982	1970	2008
	24	1895	2046	1965	2046	2074	1983	1965	2064	1969	1983	1923	2018
	25	2083	1945	1994	1965	1946	1951	1960	1952	1982	2068	1961	1981
	26	2058	2003	1968	1918	1923	1922	1916	2005	2095	2015	1973	1977
	27	1951	1920	2019	2010	2018	2044	2076	1914	1952	1933	2023	2006
	28	1964	1944	2057	2014	1937	1937	1921	1990	1951	1934	1917	1969
	29	1996	1980	2001	1980	2009	1996	1951	1911	1943	2017	1974	1945
	30	2031	1949	1966	1932	2032	2009	1994	1976	1929	1956	2053	2017
Mean # tetra-loops:		**1980.43**	**1979.10**	**1991.03**	**1991.57**	**1974.63**	**1988.63**	**1986.23**	**1978.40**	**1977.73**	**1981.00**	**1968.63**	**1981.67**
Average of means:		1981.59											
Std Dev. Of means:		6.80		Prob(obtaining ≤ 1981.589) = 0.74									
Pop'n Variance:		3.86		P Value = 0.26									
E(tetra-loops):		1977.32 ± 1.96											
Z value:		0.63											

Figure A.1: These results document the number of tetra-loops found along random sequences 10^5 nucleotides long using the basic stem-loop search algorithm.

A.1 Test of Initial Stem-loop Search Results

Refer to the spreadsheet presented in Figure A.1.

A.2 Stem-loop Probabilities

$P(G+C)$	$P(A+U)$	**P(tetra-loop)**
0.00	1.00	0.0625
0.05	0.95	0.0514
0.10	0.90	0.0429
0.15	0.85	0.0362
0.20	0.80	0.0311
0.25	0.75	0.0272
0.30	0.70	0.0243
0.35	0.65	0.0223
0.40	0.60	0.0209
0.45	0.55	0.0200
0.50	0.50	0.0198
0.55	0.45	0.0200
0.60	0.40	0.0209
0.65	0.35	0.0223
0.70	0.30	0.0243
0.75	0.25	0.0272
0.80	0.20	0.0311
0.85	0.15	0.0362
0.90	0.10	0.0429
0.95	0.05	0.0514
1.00	0.00	0.0625

Table A.1: These probabilites were calculated for tetra-loops with 4 base pairs in the stem. No unpaired or mismatched nucleotides were permitted. These values where calculated by using a short program implemented in Python. The algorithm is described in Section 1.4.2. The goal is to convey the shifting stem-loop probabilities which emerge as G+C content levels drift. This data is plotted in Figure 1.6.

A.3 Local G+C Content Levels Across Genomic Domains

Accession Number	Global G+C	CDS G+C	NC G+C	rRNA G+C	tRNA G+C
NC_002162	0.25	0.26 ± 0.05	0.18 ± 0.06	0.45 ± 0.04	0.53 ± 0.16
NC_002528	0.26	0.27 ± 0.05	0.16 ± 0.07	0.48 ± 0.06	0.54 ± 0.23
NC_001318	0.29	0.29 ± 0.05	0.22 ± 0.05	0.45 ± 0.04	0.43 ± 0.09
NC_003366	0.29	0.29 ± 0.04	0.21 ± 0.05	0.52 ± 0.04	0.56 ± 0.14
NC_004557	0.29	0.29 ± 0.04	0.25 ± 0.05	0.51 ± 0.04	0.56 ± 0.15
NC_000909	0.31	0.32 ± 0.05	0.25 ± 0.07	0.64 ± 0.03	0.63 ± 0.19
NC_003030	0.31	0.32 ± 0.04	0.25 ± 0.05	0.50 ± 0.05	0.55 ± 0.12
NC_003103	0.32	0.33 ± 0.05	0.30 ± 0.06	0.49 ± 0.05	0.56 ± 0.19
NC_003106	0.33	0.34 ± 0.04	0.29 ± 0.06	0.64 ± 0.04	0.69 ± 0.28
NC_003923	0.33	0.34 ± 0.04	0.28 ± 0.05	0.51 ± 0.04	0.59 ± 0.14
NC_005791	0.33	0.34 ± 0.05	0.23 ± 0.07	0.54 ± 0.05	0.60 ± 0.21
NC_003997	0.35	0.36 ± 0.04	0.31 ± 0.05	0.53 ± 0.04	0.59 ± 0.10
NC_004722	0.35	0.36 ± 0.04	0.31 ± 0.05	0.53 ± 0.04	0.59 ± 0.10
NC_003909	0.36	0.36 ± 0.05	0.33 ± 0.05	0.52 ± 0.04	0.52 ± 0.07
NC_004368	0.36	0.36 ± 0.04	0.29 ± 0.05	0.50 ± 0.04	0.56 ± 0.10
NC_003212	0.37	0.38 ± 0.04	0.32 ± 0.05	0.53 ± 0.04	0.58 ± 0.12
NC_004350	0.37	0.37 ± 0.05	0.31 ± 0.05	0.52 ± 0.03	0.56 ± 0.11
NC_000907	0.38	0.39 ± 0.04	0.34 ± 0.05	0.50 ± 0.04	0.47 ± 0.07
NC_002940	0.38	0.39 ± 0.04	0.33 ± 0.06	0.51 ± 0.04	0.58 ± 0.16
NC_003210	0.38	0.38 ± 0.04	0.32 ± 0.05	0.53 ± 0.04	0.58 ± 0.12
NC_003869	0.38	0.38 ± 0.05	0.34 ± 0.06	0.59 ± 0.03	0.60 ± 0.14
NC_004668	0.38	0.38 ± 0.05	0.33 ± 0.06	0.52 ± 0.04	0.55 ± 0.11
NC_002737	0.39	0.39 ± 0.04	0.33 ± 0.05	0.51 ± 0.04	0.56 ± 0.11
NC_000912	0.40	0.41 ± 0.05	0.33 ± 0.08	0.46 ± 0.03	0.52 ± 0.10
NC_002620	0.40	0.41 ± 0.03	0.36 ± 0.05	0.49 ± 0.04	0.51 ± 0.09
NC_002689	0.40	0.41 ± 0.05	0.32 ± 0.06	0.54 ± 0.04	0.60 ± 0.19
NC_000922	0.41	0.41 ± 0.04	0.34 ± 0.06	0.47 ± 0.05	0.55 ± 0.14

Continues on next page.

Local G+C Content continued...

Accession Number	Global G+C	CDS G+C	NC G+C	rRNA G+C	tRNA G+C
NC_002179	0.41	0.41 ± 0.04	0.35 ± 0.06	0.47 ± 0.04	0.46 ± 0.06
NC_003901	0.41	0.44 ± 0.06	0.33 ± 0.06	0.54 ± 0.04	0.60 ± 0.18
NC_005043	0.41	0.41 ± 0.04	0.33 ± 0.06	0.47 ± 0.05	0.54 ± 0.13
NC_000918	0.43	0.44 ± 0.05	0.37 ± 0.06	0.65 ± 0.03	0.68 ± 0.19
NC_004663	0.43	0.44 ± 0.06	0.33 ± 0.06	0.50 ± 0.03	0.55 ± 0.13
NC_000964	0.44	0.44 ± 0.05	0.37 ± 0.06	0.54 ± 0.03	0.58 ± 0.10
NC_000853	0.46	0.46 ± 0.04	0.42 ± 0.06	0.63 ± 0.03	0.65 ± 0.14
NC_003143	0.48	0.49 ± 0.06	0.41 ± 0.07	0.51 ± 0.05	0.58 ± 0.11
NC_004088	0.48	0.49 ± 0.06	0.41 ± 0.07	0.53 ± 0.02	0.58 ± 0.11
NC_000917	0.49	0.49 ± 0.05	0.39 ± 0.07	0.62 ± 0.11	0.68 ± 0.19
NC_000916	0.50	0.51 ± 0.05	0.38 ± 0.07	0.57 ± 0.03	0.62 ± 0.13
NC_004431	0.50	0.51 ± 0.06	0.46 ± 0.08	0.53 ± 0.03	0.48 ± 0.05
NC_000913	0.51	0.52 ± 0.05	0.43 ± 0.07	0.54 ± 0.03	0.59 ± 0.09
NC_002695	0.51	0.52 ± 0.06	0.43 ± 0.08	0.54 ± 0.03	0.59 ± 0.09
NC_004741	0.51	0.52 ± 0.05	0.46 ± 0.07	0.54 ± 0.03	0.54 ± 0.04
NC_003197	0.52	0.53 ± 0.06	0.44 ± 0.08	0.54 ± 0.03	0.58 ± 0.08
NC_003198	0.52	0.53 ± 0.06	0.43 ± 0.07	0.54 ± 0.03	0.59 ± 0.09
NC_004556	0.52	0.53 ± 0.05	0.47 ± 0.08	0.53 ± 0.03	0.60 ± 0.11
NC_000919	0.53	0.53 ± 0.05	0.54 ± 0.05	0.53 ± 0.03	0.57 ± 0.05
NC_002488	0.53	0.54 ± 0.06	0.47 ± 0.08	0.53 ± 0.03	0.59 ± 0.11
NC_002935	0.53	0.54 ± 0.05	0.47 ± 0.05	0.54 ± 0.03	0.59 ± 0.08
NC_002932	0.57	0.58 ± 0.06	0.47 ± 0.07	0.52 ± 0.03	0.56 ± 0.07
NC_004307	0.60	0.61 ± 0.05	0.55 ± 0.06	0.60 ± 0.03	0.59 ± 0.06
NC_002939	0.61	0.62 ± 0.06	0.55 ± 0.07	0.55 ± 0.03	0.60 ± 0.08
NC_004369	0.63	0.64 ± 0.05	0.57 ± 0.06	0.55 ± 0.03	0.60 ± 0.05
NC_003919	0.65	0.65 ± 0.04	0.62 ± 0.06	0.54 ± 0.02	0.61 ± 0.05
NC_005085	0.65	0.66 ± 0.06	0.58 ± 0.08	0.54 ± 0.02	0.59 ± 0.06
NC_002696	0.67	0.68 ± 0.04	0.62 ± 0.05	0.55 ± 0.03	0.61 ± 0.03

Continues on next page.

Local G+C Content continued...

Accession Number	**Global G+C**	**CDS G+C**	**NC G+C**	**rRNA G+C**	**tRNA G+C**
NC_002927	0.68	0.69 ± 0.05	0.63 ± 0.07	0.54 ± 0.03	0.60 ± 0.05
NC_002928	0.68	0.69 ± 0.05	0.63 ± 0.07	0.54 ± 0.03	0.60 ± 0.04
NC_002929	0.68	0.68 ± 0.05	0.62 ± 0.07	0.54 ± 0.03	0.60± 0.05

Table A.2: This table presents the average Local G+C content values and their standard deviations in the various genomic domains of the genomes in our test set. The values in this table are graphed in Figure 5.1 on page 58.

A.4 Average *bps* Across Genomic Domains

Accession Number	**Global G+C**	**CDS** *bps*	**NC** *bps*	**rRNA** *bps*	**tRNA** *bps*
NC_002162	0.25	6.04 ± 2.05	6.99 ± 2.78	7.20 ± 3.06	8.94 ± 4.92
NC_002528	0.26	6.06 ± 2.11	6.69 ± 2.58	8.91 ± 4.47	8.07 ± 3.72
NC_001318	0.29	6.17 ± 2.21	6.58 ± 3.06	7.88 ± 3.52	8.25 ± 3.99
NC_003366	0.29	6.11 ± 2.48	7.18 ± 3.19	9.00 ± 4.45	9.18 ± 4.72
NC_004557	0.29	6.11 ± 2.21	7.13 ± 3.42	9.19 ± 4.99	8.87 ± 5.33
NC_000909	0.31	6.26 ± 2.37	7.57 ± 3.87	13.1 ± 8.36	10.95 ± 5.69
NC_003030	0.31	6.19 ± 2.25	6.97 ± 3.26	8.62 ± 4.19	8.71 ± 4.41
NC_003103	0.32	6.66 ± 2.78	7.34 ± 3.85	8.83 ± 5.04	8.78 ± 4.89
NC_003106	0.33	6.45 ± 2.61	6.72 ± 3.38	13.2 ± 8.66	10.28 ± 4.86
NC_003923	0.33	6.51 ± 2.52	7.56 ± 4.00	8.58 ± 4.23	9.40 ± 6.05
NC_005791	0.33	6.57 ± 2.60	6.28 ± 2.64	10.38 ± 6.17	9.88 ± 4.90
NC_003997	0.35	6.80 ± 2.89	7.28 ± 3.57	8.92 ± 4.84	9.14 ± 5.15
NC_004722	0.35	6.84 ± 3.49	7.39 ± 3.75	8.84 ± 4.80	8.99 ± 4.79
NC_003909	0.36	7.09 ± 5.45	7.30 ± 6.01	8.94 ± 4.88	8.91 ± 4.99
NC_004368	0.36	6.75 ± 2.80	7.08 ± 3.16	8.80 ± 4.77	8.04 ± 4.42
NC_003212	0.37	7.08 ± 3.14	7.49 ± 3.62	9.23 ± 4.87	9.32 ± 5.17

Continues on next page.

Average *bps* continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	*bps*	*bps*	*bps*	*bps*
NC_004350	0.37	7.03 ± 3.12	7.03 ± 3.29	8.94 ± 4.75	8.52 ± 4.10
NC_000907	0.38	7.28 ± 3.31	7.35 ± 3.57	8.82 ± 4.61	8.89 ± 4.37
NC_002940	0.38	7.31 ± 3.48	7.44 ± 3.62	8.37 ± 4.01	9.19 ± 4.90
NC_003210	0.38	7.20 ± 3.31	7.54 ± 3.74	9.24 ± 4.90	9.37 ± 5.26
NC_003869	0.38	7.01 ± 3.37	7.82 ± 4.24	10.97 ± 6.66	9.90 ± 5.16
NC_004668	0.38	7.27 ± 3.46	7.86 ± 4.36	9.06 ± 4.80	8.26 ± 4.20
NC_002737	0.39	7.22 ± 3.30	7.12 ± 3.38	8.73 ± 4.70	8.41 ± 4.52
NC_000912	0.40	7.98 ± 4.48	8.13 ± 5.15	7.24 ± 3.11	8.38 ± 4.12
NC_002620	0.40	7.15 ± 3.20	7.18 ± 3.36	8.03 ± 3.80	9.40 ± 5.06
NC_002689	0.40	7.40 ± 3.46	6.78 ± 3.21	10.07 ± 6.37	10.10 ± 5.38
NC_000922	0.41	7.36 ± 3.42	7.04 ± 3.19	7.93 ± 3.65	8.29 ± 4.13
NC_002179	0.41	7.34 ± 3.42	7.22 ± 3.20	8.89 ± 4.77	8.22 ± 3.84
NC_003901	0.41	8.28 ± 4.42	7.17 ± 3.44	10.52 ± 6.10	10.15 ± 5.17
NC_005043	0.41	7.35 ± 3.41	7.02 ± 3.23	7.88 ± 3.62	8.29 ± 4.12
NC_000918	0.43	7.77 ± 3.87	7.43 ± 3.76	13.56 ± 9.12	11.32 ± 6.77
NC_004663	0.43	8.37 ± 4.65	8.32 ± 5.15	8.75 ± 4.40	8.11 ± 3.78
NC_000964	0.44	8.45 ± 4.57	7.78 ± 3.90	9.27 ± 5.18	9.00 ± 4.41
NC_000853	0.46	8.18 ± 4.33	8.03 ± 4.31	11.99 ± 7.48	10.91 ± 6.40
NC_003143	0.48	9.68 ± 6.02	8.22 ± 4.49	9.49 ± 5.32	9.09 ± 4.94
NC_004088	0.48	9.67 ± 6.07	8.46 ± 4.73	9.73 ± 5.36	8.72 ± 4.92
NC_000917	0.49	8.86 ± 5.01	8.60 ± 5.13	10.67 ± 7.65	11.02 ± 4.86
NC_000916	0.50	9.72 ± 5.94	7.98 ± 4.55	9.58 ± 5.58	11.19 ± 6.73
NC_004431	0.50	10.61 ± 6.89	9.87 ± 6.43	9.29 ± 4.92	8.58 ± 4.27
NC_000913	0.51	10.76 ± 7.06	9.32 ± 6.27	9.31 ± 4.85	9.27 ± 5.04
NC_002695	0.51	10.78 ± 7.13	9.34 ± 6.31	9.27 ± 4.68	9.48 ± 5.44
NC_004741	0.51	10.69 ± 6.91	9.73 ± 6.15	9.63 ± 5.62	9.87 ± 5.79
NC_003197	0.52	11.40 ± 7.81	9.70 ± 6.52	9.43 ± 4.78	9.06 ± 4.68
NC_003198	0.52	11.29 ± 7.67	9.09 ± 5.41	9.68 ± 5.34	9.37 ± 4.95

Continues on next page.

Average *bps* continued...

Accession **Number**	**Global** **G+C**	**CDS** *bps*	**NC** *bps*	**rRNA** *bps*	**tRNA** *bps*
NC_004556	0.52	11.06 ± 7.75	9.33 ± 6.02	8.72 ± 4.77	9.70 ± 5.28
NC_000919	0.53	10.52 ± 7.08	10.9 ± 7.47	8.46 ± 3.91	12.33 ± 9.31
NC_002488	0.53	11.36 ± 8.74	9.75 ± 6.94	8.66 ± 4.74	9.72 ± 5.30
NC_002935	0.53	11.43 ± 8.17	9.81 ± 6.46	9.88 ± 6.06	10.29 ± 6.23
NC_002932	0.57	13.75 ± 10.34	10.52 ± 7.31	9.16 ± 4.96	9.25 ± 5.27
NC_004307	0.60	15.40 ± 12.15	13.29 ± 10.62	11.14 ± 7.10	10.06 ± 6.21
NC_002939	0.61	14.06 ± 10.38	11.88 ± 8.58	9.32 ± 4.81	9.92 ± 5.54
NC_004369	0.63	15.69 ± 12.11	12.69 ± 9.13	9.46 ± 5.73	11.33 ± 7.15
NC_003919	0.65	18.41 ± 14.99	16.68 ± 13.51	10.34 ± 6.23	10.54 ± 6.11
NC_005085	0.65	17.83 ± 14.65	14.30 ± 11.40	9.15 ± 4.82	10.16 ± 5.65
NC_002696	0.67	18.55 ± 14.69	15.11 ± 11.91	10.64 ± 7.17	11.81 ± 8.67
NC_002927	0.68	19.40 ± 15.96	15.19 ± 11.95	9.70 ± 5.39	9.85 ± 6.16
NC_002928	0.68	19.32 ± 15.88	15.19 ± 11.89	9.38 ± 5.06	9.97 ± 6.20
NC_002929	0.68	19.41 ± 15.88	15.01 ± 11.49	9.69 ± 5.36	10.79 ± 7.32

Table A.3: This table presents the average *bps* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.2 and B.1 on pages 59 and 144, respectively.

A.5 Average *span* Across Genomic Domains

Accession Number	**Global G+C**	**CDS** *span*	**NC** *span*	**rRNA** *span*	**tRNA** *span*
NC_002162	0.25	22.27 ± 7.01	23.5 ± 7.48	26.35 ± 10.22	31.47 ± 15.21
NC_002528	0.26	22.47 ± 7.3	23.28 ± 7.34	31.66 ± 13.89	29.24 ± 12.38
NC_001318	0.29	22.84 ± 7.58	23.97 ± 9.72	28.32 ± 11.79	29.82 ± 13.37
NC_003366	0.29	22.63 ± 8.46	24.21 ± 8.4	32.02 ± 14.77	31.98 ± 14.90
NC_004557	0.29	22.65 ± 7.53	24.38 ± 9.17	32.45 ± 16.03	31.25 ± 16.5
NC_000909	0.31	23.23 ± 7.98	26.52 ± 11.92	44.98 ± 26.81	38.03 ± 18.07
NC_003030	0.31	23.03 ± 7.65	23.98 ± 8.69	30.83 ± 13.48	31.19 ± 14.00
NC_003103	0.32	24.69 ± 9.14	26.12 ± 11.42	31.74 ± 16.23	31.3 ± 16.15
NC_003106	0.33	23.81 ± 8.67	24.46 ± 10.57	45.35 ± 27.44	35.89 ± 15.29
NC_003923	0.33	23.96 ± 8.42	25.97 ± 11.42	30.69 ± 13.71	33.17 ± 19.68
NC_005791	0.33	24.21 ± 8.63	22.77 ± 8.72	36.13 ± 18.99	34.22 ± 15.04
NC_003997	0.35	24.97 ± 9.45	25.11 ± 9.59	32.07 ± 16.08	32.35 ± 16.86
NC_004722	0.35	25.11 ± 11.71	25.52 ± 10.29	31.76 ± 15.94	31.88 ± 15.62
NC_003909	0.36	25.69 ± 15.80	25.95 ± 16.75	32.1 ± 16.21	31.76 ± 16.42
NC_004368	0.36	24.72 ± 9.17	24.34 ± 8.58	31.61 ± 15.89	29.36 ± 14.92
NC_003212	0.37	26.15 ± 10.29	25.70 ± 9.53	32.76 ± 15.75	32.79 ± 16.37
NC_004350	0.37	25.71 ± 10.13	24.75 ± 9.64	31.85 ± 15.23	30.54 ± 13.61
NC_000907	0.38	26.50 ± 10.72	26.04 ± 10.85	31.68 ± 14.83	31.28 ± 14.39
NC_002940	0.38	26.67 ± 11.33	26.35 ± 11.15	30.23 ± 13.01	32.37 ± 15.95
NC_003210	0.38	26.51 ± 10.88	25.70 ± 9.69	32.78 ± 15.83	32.97 ± 16.64
NC_003869	0.38	25.67 ± 10.88	27.39 ± 12.90	38.37 ± 21.54	34.59 ± 16.50
NC_004668	0.38	26.47 ± 11.22	26.84 ± 11.83	32.36 ± 15.87	29.5 ± 14.00
NC_002737	0.39	26.21 ± 10.61	25.16 ± 10.04	31.08 ± 15.42	30.57 ± 15.20
NC_000912	0.40	28.58 ± 14.20	28.74 ± 16.09	26.66 ± 10.10	29.57 ± 13.56
NC_002620	0.40	25.91 ± 10.31	25.45 ± 10.15	29.43 ± 12.72	32.96 ± 15.72
NC_002689	0.40	27.05 ± 11.26	24.91 ± 10.43	35.2 ± 19.16	35.14 ± 17.62
NC_000922	0.41	26.61 ± 10.94	25.18 ± 9.68	28.66 ± 11.96	29.33 ± 12.91

Continues on next page.

Average *span* continued...

Accession **Number**	**Global** **G+C**	**CDS** *span*	**NC** *span*	**rRNA** *span*	**tRNA** *span*
NC_002179	0.41	26.53 ± 10.94	25.74 ± 9.86	31.79 ± 15.89	28.98 ± 12.32
NC_003901	0.41	29.84 ± 14.16	25.98 ± 10.91	36.89 ± 19.57	35.09 ± 16.06
NC_005043	0.41	26.59 ± 10.91	25.03 ± 9.75	28.46 ± 11.83	29.32 ± 12.89
NC_000918	0.43	28.07 ± 12.44	26.92 ± 11.89	46.94 ± 29.28	38.99 ± 20.84
NC_004663	0.43	30.10 ± 14.84	28.39 ± 14.27	31.15 ± 13.57	28.97 ± 12.43
NC_000964	0.44	30.45 ± 14.68	27.16 ± 11.53	33.10 ± 17.03	31.90 ± 14.58
NC_000853	0.46	29.31 ± 13.75	28.43 ± 13.64	41.73 ± 24.49	37.71 ± 20.49
NC_003143	0.48	34.21 ± 19.01	29.22 ± 13.95	33.82 ± 16.73	32.14 ± 16.12
NC_004088	0.48	34.18 ± 19.16	30.03 ± 14.77	34.56 ± 16.61	30.81 ± 16.00
NC_000917	0.49	31.70 ± 15.87	30.78 ± 16.34	36.60 ± 25.42	38.53 ± 17.28
NC_000916	0.50	34.27 ± 18.74	28.71 ± 14.50	34.03 ± 17.37	38.63 ± 20.95
NC_004431	0.50	37.15 ± 21.76	34.69 ± 20.14	33.18 ± 15.80	30.82 ± 14.23
NC_000913	0.51	37.63 ± 22.31	32.39 ± 19.08	33.13 ± 15.47	33.14 ± 16.83
NC_002695	0.51	37.71 ± 22.52	32.69 ± 19.47	33.09 ± 15.09	33.85 ± 17.86
NC_004741	0.51	37.43 ± 21.85	34.20 ± 19.40	34.30 ± 17.75	34.53 ± 18.83
NC_003197	0.52	39.75 ± 24.70	33.89 ± 20.41	33.53 ± 15.48	32.47 ± 15.67
NC_003198	0.52	39.42 ± 24.32	31.88 ± 16.96	34.44 ± 17.36	33.30 ± 16.40
NC_004556	0.52	38.37 ± 24.09	32.86 ± 18.83	31.37 ± 15.82	34.19 ± 17.65
NC_000919	0.53	36.57 ± 22.05	37.73 ± 23.33	30.72 ± 13.21	42.37 ± 29.62
NC_002488	0.53	39.47 ± 27.23	34.28 ± 21.62	31.17 ± 15.76	34.21 ± 17.51
NC_002935	0.53	39.52 ± 25.45	33.80 ± 19.43	35.03 ± 19.06	36.02 ± 20.29
NC_002932	0.57	47.09 ± 32.66	36.72 ± 23.13	32.33 ± 15.48	32.14 ± 16.85
NC_004307	0.60	52.16 ± 38.08	45.16 ± 33.06	39.66 ± 23.39	35.51 ± 20.69
NC_002939	0.61	48.25 ± 32.97	41.10 ± 27.12	33.31 ± 15.71	34.59 ± 18.18
NC_004369	0.63	53.04 ± 37.99	43.16 ± 28.20	33.55 ± 18.06	38.38 ± 22.5
NC_003919	0.65	61.27 ± 46.70	55.84 ± 42.15	36.27 ± 19.80	37.07 ± 19.78
NC_005085	0.65	59.48 ± 45.61	48.51 ± 35.81	32.48 ± 15.44	35.84 ± 18.45
NC_002696	0.67	62.01 ± 46.13	50.99 ± 37.17	37.44 ± 22.92	41.01 ± 27.50

Continues on next page.

Average *span* continued...

Accession Number	**Global G+C**	**CDS** *span*	**NC** *span*	**rRNA** *span*	**tRNA** *span*
NC_002927	0.68	64.28 ± 49.67	51.09 ± 37.47	34.43 ± 17.15	34.28 ± 19.77
NC_002928	0.68	64.06 ± 49.46	51.10 ± 37.28	33.50 ± 16.19	34.93 ± 20.43
NC_002929	0.68	64.37 ± 49.52	50.52 ± 36.01	34.38 ± 16.98	37.64 ± 23.86

Table A.4: This table presents the average *span* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.4 and B.3 on pages 63 and 147, respectively.

A.6 Average *cSpacing* Across Genomic Domains

Accession Number	**Global G+C**	**CDS** *cSpacing*	**NC** *cSpacing*	**rRNA** *cSpacing*	**tRNA** *cSpacing*
NC_002162	0.25	140.24 ± 103.78	161.60 ± 113.84	49.12 ± 21.30	74.68 ± 62.10
NC_002528	0.26	121.30 ± 87.14	162.80 ± 107.30	46.44 ± 13.46	120.93 ± 79.43
NC_001318	0.29	111.04 ± 72.92	134.73 ± 107.06	47.94 ± 20.63	89.45 ± 63.12
NC_003366	0.29	107.86 ± 78.65	149.72 ± 103.90	44.98 ± 14.11	70.01 ± 77.23
NC_004557	0.29	109.30 ± 75.49	124.38 ± 85.08	45.13 ± 15.14	63.46 ± 59.09
NC_000909	0.31	95.32 ± 60.46	99.87 ± 71.63	44.87 ± 14.61	93.65 ± 71.59
NC_003030	0.31	95.11 ± 61.70	122.88 ± 75.92	44.65 ± 14.20	49.02 ± 30.23
NC_003103	0.32	79.15 ± 46.21	86.38 ± 51.38	47.60 ± 14.84	74.47 ± 48.94
NC_003106	0.33	85.40 ± 51.48	96.39 ± 61.06	46.82 ± 17.80	97.31 ± 75.76
NC_003923	0.33	79.14 ± 46.40	99.55 ± 63.41	43.08 ± 13.75	41.08 ± 18.32
NC_005791	0.33	75.71 ± 44.33	105.97 ± 60.77	45.67 ± 13.37	81.1 ± 54.33
NC_003997	0.35	70.46 ± 38.67	86.42 ± 50.01	41.76 ± 14.22	45.54 ± 20.76
NC_004722	0.35	70.93 ± 39.52	85.25 ± 49.77	41.21 ± 12.92	45.92 ± 24.45
NC_003909	0.36	71.08 ± 39.34	79.18 ± 48.59	41.95 ± 13.76	45.96 ± 21.51

Continues on next page.

Average *cSpacing* continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	*cSpacing*	*cSpacing*	*cSpacing*	*cSpacing*
NC_004368	0.36	68.18 ± 36.56	85.56 ± 48.68	43.49 ± 13.84	45.99 ± 21.51
NC_003212	0.37	64.45 ± 31.52	82.78 ± 41.55	42.70 ± 11.92	51.19 ± 24.70
NC_004350	0.37	63.30 ± 32.95	78.62 ± 43.73	42.65 ± 12.98	45.96 ± 20.66
NC_000907	0.38	58.24 ± 26.42	67.88 ± 34.24	46.08 ± 15.88	57.23 ± 31.02
NC_002940	0.38	58.67 ± 26.98	65.95 ± 32.45	46.15 ± 16.82	59.57 ± 34.22
NC_003210	0.38	62.79 ± 30.46	82.55 ± 42.20	42.76 ± 12.02	49.41 ± 23.13
NC_003869	0.38	66.02 ± 35.20	72.25 ± 41.92	43.48 ± 14.09	47.31 ± 22.88
NC_004668	0.38	61.62 ± 30.49	75.19 ± 41.60	44.32 ± 15.72	45.66 ± 23.12
NC_002737	0.39	57.76 ± 28.15	69.03 ± 35.72	43.39 ± 13.69	48.60 ± 29.76
NC_000912	0.40	54.97 ± 23.46	65.17 ± 34.92	48.25 ± 19.22	43.06 ± 14.05
NC_002620	0.40	55.32 ± 22.74	63.72 ± 28.44	45.63 ± 14.78	53.22 ± 25.80
NC_002689	0.40	57.07 ± 25.77	70.75 ± 40.27	43.16 ± 14.64	64.15 ± 36.77
NC_000922	0.41	55.11 ± 23.34	67.13 ± 32.74	48.46 ± 18.61	49.84 ± 21.61
NC_002179	0.41	54.94 ± 23.12	66.46 ± 31.43	46.26 ± 16.30	51.95 ± 21.08
NC_003901	0.41	50.57 ± 21.84	72.81 ± 41.78	46.10 ± 17.47	60.79 ± 34.87
NC_005043	0.41	55.14 ± 23.44	68.65 ± 33.37	48.14 ± 18.45	50.14 ± 21.65
NC_000918	0.43	52.48 ± 22.15	60.98 ± 28.83	45.85 ± 16.07	62.61 ± 29.68
NC_004663	0.43	51.67 ± 21.91	69.60 ± 34.75	44.51 ± 12.67	50.49 ± 24.51
NC_000964	0.44	51.11 ± 20.66	64.45 ± 30.98	41.10 ± 11.99	43.79 ± 15.82
NC_000853	0.46	48.05 ± 17.60	53.50 ± 23.84	43.26 ± 13.40	49.07 ± 21.91
NC_003143	0.48	47.13 ± 16.62	54.36 ± 24.70	45.90 ± 14.66	44.25 ± 16.53
NC_004088	0.48	47.12 ± 16.51	53.16 ± 23.90	46.58 ± 15.69	44.36 ± 17.85
NC_000917	0.49	48.50 ± 17.65	54.24 ± 22.79	42.12 ± 17.57	54.48 ± 20.00
NC_000916	0.50	46.36 ± 16.19	55.80 ± 24.61	43.71 ± 13.95	53.39 ± 21.45
NC_004431	0.50	46.52 ± 15.67	49.03 ± 19.48	44.04 ± 12.80	45.81 ± 15.51
NC_000913	0.51	46.37 ± 15.54	51.80 ± 21.33	44.37 ± 12.84	46.64 ± 17.85
NC_002695	0.51	46.53 ± 16.01	51.60 ± 21.69	44.41 ± 12.70	45.89 ± 16.98
NC_004741	0.51	46.11 ± 15.22	48.83 ± 18.54	43.96 ± 13.61	43.73 ± 17.11

Continues on next page.

Average *cSpacing* continued...

Accession **Number**	**Global** **G+C**	**CDS** *cSpacing*	**NC** *cSpacing*	**rRNA** *cSpacing*	**tRNA** *cSpacing*
NC_003197	0.52	46.48 ± 15.91	50.36 ± 20.07	44.98 ± 13.00	43.24 ± 14.32
NC_003198	0.52	46.42 ± 15.86	50.56 ± 20.09	43.69 ± 12.54	44.22 ± 16.98
NC_004556	0.52	48.02 ± 16.95	52.33 ± 22.15	42.71 ± 12.73	47.54 ± 18.99
NC_000919	0.53	46.50 ± 15.67	47.50 ± 16.16	43.29 ± 12.82	47.29 ± 19.37
NC_002488	0.53	48.15 ± 17.86	51.92 ± 22.30	43.21 ± 12.63	45.69 ± 16.26
NC_002935	0.53	46.75 ± 16.05	48.59 ± 16.57	46.18 ± 15.24	42.08 ± 14.42
NC_002932	0.57	48.95 ± 18.63	49.69 ± 18.17	44.21 ± 13.97	43.99 ± 15.76
NC_004307	0.60	50.60 ± 20.58	49.12 ± 18.95	44.67 ± 14.17	41.65 ± 16.03
NC_002939	0.61	48.24 ± 18.22	48.32 ± 17.29	43.86 ± 12.23	43.66 ± 14.86
NC_004369	0.63	50.33 ± 20.36	48.11 ± 16.96	42.42 ± 12.84	42.67 ± 14.52
NC_003919	0.65	54.21 ± 24.44	52.38 ± 22.41	43.52 ± 12.49	41.86 ± 13.30
NC_005085	0.65	53.43 ± 23.96	50.47 ± 20.07	43.79 ± 12.88	43.79 ± 14.21
NC_002696	0.67	54.27 ± 24.07	49.72 ± 20.06	43.40 ± 13.49	43.16 ± 16.48
NC_002927	0.68	55.50 ± 25.83	50.74 ± 20.48	42.63 ± 11.78	41.72 ± 13.19
NC_002928	0.68	55.38 ± 25.74	50.69 ± 20.32	42.70 ± 12.28	42.11 ± 13.66
NC_002929	0.68	55.65 ± 25.75	50.81 ± 19.78	42.56 ± 11.71	43.56 ± 15.21

Table A.5: This table presents the average *cSpacing* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.6, 5.7 and B.5 on pages 67, 68, and 149, respectively.

A.7 Average *fSpacing* Across Genomic Domains

Accession Number	Global G+C	CDS *fSpacing*	NC *fSpacing*	rRNA *fSpacing*	tRNA *fSpacing*
NC_002162	0.25	119.47 ± 104.12	139.99 ± 114.06	25.38 ± 20.68	48.09 ± 60.75
NC_002528	0.26	100.41 ± 87.34	139.35 ± 103.34	18.91 ± 11.91	96.20 ± 77.85
NC_001318	0.29	89.88 ± 73.08	112.74 ± 107.37	22.74 ± 19.90	65.07 ± 62.29
NC_003366	0.29	86.83 ± 78.78	127.76 ± 104.26	17.64 ± 11.78	43.59 ± 76.59
NC_004557	0.29	88.23 ± 75.64	102.33 ± 85.22	17.45 ± 13.66	37.30 ± 58.31
NC_000909	0.31	73.83 ± 60.52	76.52 ± 72.56	10.78 ± 5.71	63.48 ± 70.53
NC_003030	0.31	73.72 ± 61.70	101.03 ± 76.01	17.95 ± 11.50	23.03 ± 29.55
NC_003103	0.32	56.51 ± 46.24	62.91 ± 51.46	20.55 ± 13.66	48.70 ± 46.67
NC_003106	0.33	63.47 ± 51.58	74.09 ± 61.24	12.16 ± 10.84	68.84 ± 75.05
NC_003923	0.33	57.02 ± 46.33	76.28 ± 64.00	16.64 ± 11.64	13.60 ± 9.51
NC_005791	0.33	53.52 ± 44.23	84.80 ± 60.98	15.57 ± 9.17	53.33 ± 53.03
NC_003997	0.35	47.59 ± 38.58	63.55 ± 50.20	14.58 ± 11.27	18.40 ± 17.92
NC_004722	0.35	48.00 ± 39.41	62.10 ± 49.86	14.23 ± 9.69	19.03 ± 22.62
NC_003909	0.36	47.84 ± 39.02	55.78 ± 48.51	14.71 ± 10.78	19.22 ± 19.56
NC_004368	0.36	45.50 ± 36.44	63.23 ± 48.83	17.11 ± 11.39	20.68 ± 20.71
NC_003212	0.37	40.69 ± 31.24	59.42 ± 41.59	14.87 ± 9.15	23.52 ± 22.37
NC_004350	0.37	39.98 ± 32.82	56.02 ± 43.69	15.66 ± 10.52	20.21 ± 18.15
NC_000907	0.38	34.37 ± 25.99	44.34 ± 34.14	19.14 ± 13.85	31.28 ± 29.22
NC_002940	0.38	34.72 ± 26.58	42.26 ± 32.52	20.02 ± 15.13	33.10 ± 31.65
NC_003210	0.38	38.78 ± 30.13	59.28 ± 42.13	14.93 ± 9.19	21.77 ± 20.53
NC_003869	0.38	42.77 ± 35.11	48.12 ± 42.23	13.12 ± 8.24	19.41 ± 20.58
NC_004668	0.38	37.80 ± 30.34	51.23 ± 41.63	17.09 ± 12.89	20.44 ± 21.93
NC_002737	0.39	34.08 ± 27.84	46.10 ± 35.71	17.19 ± 12.13	22.85 ± 26.64
NC_000912	0.40	29.73 ± 22.94	40.10 ± 35.75	24.39 ± 18.46	17.6 ± 12.13
NC_002620	0.40	31.94 ± 22.15	40.56 ± 27.96	19.88 ± 12.76	26.58 ± 23.62
NC_002689	0.40	32.84 ± 25.22	47.89 ± 40.57	14.35 ± 9.83	36.18 ± 33.30
NC_000922	0.41	31.17 ± 22.75	43.92 ± 32.44	23.43 ± 17.05	25.15 ± 19.15

Continues on next page.

Average *fSpacing* continued...

Accession	Global	CDS	NC	rRNA	tRNA
Number	G+C	*fSpacing*	*fSpacing*	*fSpacing*	*fSpacing*
NC_002179	0.41	31.06 ± 22.37	42.87 ± 31.44	18.98 ± 14.81	27.3 ± 19.77
NC_003901	0.41	24.63 ± 21.02	49.45 ± 41.76	15.87 ± 14.74	32.64 ± 33.34
NC_005043	0.41	31.22 ± 22.86	45.50 ± 33.01	23.19 ± 16.96	25.49 ± 19.28
NC_000918	0.43	27.67 ± 21.29	36.96 ± 28.71	10.62 ± 5.04	32.56 ± 26.08
NC_004663	0.43	25.52 ± 21.10	44.59 ± 34.59	17.97 ± 11.36	25.16 ± 22.74
NC_000964	0.44	24.73 ± 19.57	40.17 ± 30.94	13.71 ± 8.93	16.69 ± 13.70
NC_000853	0.46	22.45 ± 16.23	28.48 ± 23.35	11.12 ± 5.66	19.35 ± 17.26
NC_003143	0.48	18.57 ± 13.84	28.81 ± 24.18	17.23 ± 12.02	17.47 ± 13.21
NC_004088	0.48	18.58 ± 13.71	27.08 ± 23.36	17.56 ± 12.93	18.84 ± 14.03
NC_000917	0.49	21.31 ± 15.74	27.87 ± 22.52	14.07 ± 6.95	25.59 ± 15.40
NC_000916	0.50	17.86 ± 13.33	30.53 ± 24.04	14.97 ± 10.07	23.58 ± 18.09
NC_004431	0.50	16.34 ± 11.60	20.33 ± 17.49	16.09 ± 10.09	19.21 ± 14.23
NC_000913	0.51	15.91 ± 10.96	24.42 ± 19.98	16.22 ± 9.94	18.74 ± 13.99
NC_002695	0.51	16.06 ± 11.60	24.02 ± 20.47	16.33 ± 10.10	17.92 ± 12.82
NC_004741	0.51	15.79 ± 10.58	20.35 ± 16.65	15.01 ± 10.10	16.07 ± 12.73
NC_003197	0.52	14.99 ± 10.28	22.18 ± 18.34	16.43 ± 10.69	16.04 ± 11.36
NC_003198	0.52	15.12 ± 10.46	23.48 ± 18.88	14.61 ± 8.94	16.60 ± 13.46
NC_004556	0.52	17.01 ± 12.21	24.54 ± 20.97	15.65 ± 9.47	19.35 ± 15.35
NC_000919	0.53	16.83 ± 11.36	17.18 ± 11.77	16.55 ± 11.25	15.16 ± 11.76
NC_002488	0.53	16.69 ± 12.17	23.31 ± 20.86	16.31 ± 9.36	17.82 ± 12.05
NC_002935	0.53	15.20 ± 9.85	20.14 ± 13.74	16.97 ± 11.81	12.88 ± 8.39
NC_002932	0.57	13.40 ± 8.91	19.83 ± 15.04	16.73 ± 10.99	16.73 ± 12.42
NC_004307	0.60	12.27 ± 7.17	14.56 ± 9.65	13.22 ± 7.26	13.19 ± 9.77
NC_002939	0.61	12.22 ± 7.34	16.07 ± 11.51	15.65 ± 9.10	14.92 ± 10.00
NC_004369	0.63	11.68 ± 6.60	14.65 ± 9.38	14.41 ± 8.80	12.60 ± 7.07
NC_003919	0.65	11.25 ± 5.97	12.26 ± 7.17	13.65 ± 9.06	11.70 ± 7.44
NC_005085	0.65	11.51 ± 6.55	14.32 ± 10.05	16.02 ± 10.40	14.10 ± 9.57
NC_002696	0.67	10.95 ± 5.64	12.14 ± 6.93	13.21 ± 6.87	10.84 ± 5.80

Continues on next page.

Average *fSpacing* continued...

Accession Number	Global G+C	CDS *fSpacing*	NC *fSpacing*	rRNA *fSpacing*	tRNA *fSpacing*
NC_002927	0.68	11.09 ± 5.84	13.24 ± 8.34	13.69 ± 8.02	13.40 ± 7.51
NC_002928	0.68	11.09 ± 5.87	13.20 ± 8.23	14.50 ± 9.12	13.54 ± 7.52
NC_002929	0.68	11.19 ± 5.94	13.49 ± 8.19	13.61 ± 8.06	13.51 ± 7.07

Table A.6: This table presents the average *fSpacing* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.9 and B.7 on pages 71 and 151, respectively.

A.8 Average $(cSpacing \times bps)$ Across Genomic Domains

Accession	Global	$(cSpacing \times bps)$			
Number	G+C	CDS	NC	rRNA	tRNA
NC_002162	0.25	835.12 ± 662.96	1132.59 ± 945.46	367.16 ± 242.48	736.82 ± 1005.09
NC_002528	0.26	731.45 ± 585.97	1066.61 ± 782.05	438.89 ± 318.80	1069.02 ± 1077.31
NC_001318	0.29	684.36 ± 525.65	883.62 ± 826.13	394.47 ± 283.42	775.77 ± 713.36
NC_003366	0.29	658.18 ± 581.54	1064.52 ± 856.89	433.24 ± 331.83	703.55 ± 1021.79
NC_004557	0.29	666.20 ± 524.75	898.68 ± 823.31	445.84 ± 362.16	610.50 ± 762.89
NC_000909	0.31	596.76 ± 452.83	732.2 ± 622.87	695.66 ± 739.40	1126.37 ± 1205.09
NC_003030	0.31	590.88 ± 453.12	861.71 ± 701.09	413.76 ± 321.08	462.45 ± 454.32
NC_003103	0.32	531.50 ± 396.08	643.91 ± 576.85	447.21 ± 369.33	741.79 ± 876.50
NC_003106	0.33	552.58 ± 413.26	648.91 ± 578.59	735.53 ± 809.66	1070.02 ± 1061.69
NC_003923	0.33	519.95 ± 383.56	746.69 ± 627.51	396.49 ± 302.25	479.37 ± 609.09
NC_005791	0.33	503.86 ± 375.72	657.96 ± 474.79	530.67 ± 483.19	857.77 ± 822.98
NC_003997	0.35	486.21 ± 362.83	636.80 ± 518.59	408.84 ± 347.15	467.79 ± 475.39
NC_004722	0.35	495.91 ± 590.23	639.63 ± 534.03	399.43 ± 337.57	457.60 ± 463.31
NC_003909	0.36	533.98 ± 1166.64	607.87 ± 1019.51	410.86 ± 348.25	450.80 ± 427.86
NC_004368	0.36	467.20 ± 337.12	609.91 ± 464.27	415.13 ± 343.63	400.12 ± 383.25
NC_003212	0.37	467.65 ± 343.17	624.47 ± 453.83	428.89 ± 344.77	522.75 ± 472.59

Continues on next page.

Average $\left(cSpacing \times bps\right)$ continued...

Accession	**Global**	$\left(cSpacing \times bps\right)$			
Number	**G+C**	**CDS**	**NC**	**rRNA**	**tRNA**
NC_004350	0.37	454.34 ± 338.52	563.32 ± 459.33	412.46 ± 339.07	431.41 ± 356.72
NC_000907	0.38	437.45 ± 322.32	511.88 ± 414.58	437.69 ± 362.21	559.24 ± 489.72
NC_002940	0.38	442.85 ± 340.20	502.02 ± 394.45	412.04 ± 314.06	618.04 ± 668.00
NC_003210	0.38	464.62 ± 357.30	630.43 ± 482.02	430.74 ± 348.02	519.76 ± 499.27
NC_003869	0.38	472.02 ± 389.97	577.41 ± 521.87	549.37 ± 547.25	516.96 ± 470.38
NC_004668	0.38	459.28 ± 349.74	605.89 ± 523.85	438.94 ± 376.51	412.77 ± 404.34
NC_002737	0.39	428.38 ± 318.77	499.84 ± 389.16	407.02 ± 342.39	472.92 ± 588.37
NC_000912	0.40	460.27 ± 404.32	536.28 ± 476.55	361.42 ± 239.77	388.85 ± 307.56
NC_002620	0.40	408.98 ± 289.53	472.39 ± 354.98	391.03 ± 283.18	555.18 ± 538.74
NC_002689	0.40	437.58 ± 331.71	480.56 ± 363.62	500.89 ± 522.49	736.72 ± 775.29
NC_000922	0.41	420.10 ± 308.92	482.16 ± 357.61	412.20 ± 311.30	457.61 ± 442.82
NC_002179	0.41	419.49 ± 312.97	486.48 ± 340.36	442.62 ± 391.85	456.62 ± 382.42
NC_003901	0.41	443.12 ± 375.91	530.19 ± 442.45	539.61 ± 509.97	678.14 ± 639.47
NC_005043	0.41	419.82 ± 307.92	492.13 ± 365.01	405.55 ± 305.46	459.71 ± 441.97
NC_000918	0.43	427.75 ± 337.62	465.12 ± 361.56	755.82 ± 875.66	808.70 ± 789.55
NC_004663	0.43	458.77 ± 409.92	601.36 ± 562.34	415.21 ± 305.32	445.44 ± 404.95
NC_000964	0.44	459.27 ± 394.94	510.14 ± 379.38	418.89 ± 363.88	429.05 ± 343.49
NC_000853	0.46	418.79 ± 351.25	446.28 ± 380.71	604.85 ± 611.89	623.54 ± 643.08
NC_003143	0.48	507.67 ± 568.38	468.01 ± 401.40	474.97 ± 405.45	449.00 ± 422.67
NC_004088	0.48	507.65 ± 583.13	473.79 ± 414.98	494.02 ± 419.48	438.59 ± 444.17
NC_000917	0.49	465.33 ± 416.48	489.16 ± 431.88	564.90 ± 685.80	656.94 ± 491.23
NC_000916	0.50	501.30 ± 502.62	466.27 ± 432.48	466.03 ± 459.49	673.76 ± 641.04
NC_004431	0.50	562.47 ± 623.19	537.48 ± 610.23	443.97 ± 368.25	418.28 ± 318.15
NC_000913	0.51	572.47 ± 658.40	527.95 ± 622.49	446.66 ± 362.81	483.15 ± 457.69
NC_002695	0.51	575.98 ± 669.41	527.78 ± 772.50	441.64 ± 339.47	490.96 ± 482.26
NC_004741	0.51	563.99 ± 622.02	523.73 ± 546.77	467.83 ± 449.37	495.60 ± 508.90
NC_003197	0.52	620.50 ± 762.55	540.13 ± 649.57	454.64 ± 336.20	428.45 ± 350.73
NC_003198	0.52	611.58 ± 735.66	494.45 ± 468.59	464.95 ± 402.48	462.62 ± 443.82
NC_004556	0.52	617.92 ± 779.34	530.72 ± 556.00	408.37 ± 375.42	514.90 ± 467.93
NC_000919	0.53	561.45 ± 660.15	595.82 ± 683.54	386.79 ± 260.06	718.63 ± 951.61
NC_002488	0.53	656.83 ± 1072.27	562.35 ± 724.70	408.73 ± 374.86	499.13 ± 453.64
NC_002935	0.53	632.89 ± 830.89	530.69 ± 575.54	509.31 ± 518.65	504.71 ± 520.54

Continues on next page.

Average $\left(cSpacing \times bps\right)$ continued...

Accession	Global	$\left(cSpacing \times bps\right)$			
Number	**G+C**	**CDS**	**NC**	**rRNA**	**tRNA**
NC_002932	0.57	837.17 ± 1196.41	594.27 ± 714.97	441.80 ± 368.61	457.80 ± 437.30
NC_004307	0.60	1007.46 ± 1581.57	821.15 ± 1315.83	578.73 ± 620.14	497.13 ± 550.77
NC_002939	0.61	846.21 ± 1172.83	680.18 ± 919.13	446.47 ± 392.06	491.64 ± 481.90
NC_004369	0.63	1017.06 ± 1517.56	734.11 ± 955.50	451.86 ± 472.02	568.83 ± 661.36
NC_003919	0.65	1346.74 ± 2187.19	1154.38 ± 1877.78	503.50 ± 529.94	504.34 ± 525.91
NC_005085	0.65	1284.25 ± 2229.62	914.86 ± 1409.86	434.55 ± 358.73	498.99 ± 449.62
NC_002696	0.67	1343.80 ± 2090.87	969.50 ± 1574.23	541.82 ± 634.14	637.68 ± 867.51
NC_002927	0.68	1471.28 ± 2446.71	988.25 ± 1548.71	455.48 ± 405.22	473.16 ± 506.79
NC_002928	0.68	1461.24 ± 2441.53	984.41 ± 1519.85	438.50 ± 376.06	486.18 ± 507.61
NC_002929	0.68	1471.08 ± 2400.34	963.47 ± 1402.06	453.37 ± 402.14	563.51 ± 655.32

Table A.7: This table presents the average $\left(cSpacing \times bps\right)$ values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.11 and B.9 on pages 74 and 153, respectively.

A.9 Average $(fSpacing \times bps)$ Across Genomic Domains

Accession	Global	$\left(fSpacing \times bps\right)$			
Number	**G+C**	**CDS**	**NC**	**rRNA**	**tRNA**
NC_002162	0.25	703.80 ± 643.35	973.46 ± 909.66	181.23 ± 162.98	463.64 ± 838.97
NC_002528	0.26	598.56 ± 562.68	907.36 ± 747.29	162.87 ± 130.05	848.51 ± 954.76
NC_001318	0.29	546.68 ± 493.65	725.77 ± 769.67	176.38 ± 182.06	551.11 ± 583.26
NC_003366	0.29	520.21 ± 514.83	895.72 ± 810.08	155.99 ± 131.64	426.48 ± 899.91
NC_004557	0.29	530.40 ± 496.57	728.12 ± 753.31	151.71 ± 131.15	335.92 ± 595.55
NC_000909	0.31	454.18 ± 415.82	533.89 ± 556.95	140.79 ± 123.69	745.92 ± 1038.82
NC_003030	0.31	451.16 ± 419.36	697.31 ± 639.24	156.45 ± 134.56	206.71 ± 322.39
NC_003103	0.32	369.29 ± 338.76	451.04 ± 444.74	170.38 ± 128.07	476.74 ± 678.69
NC_003106	0.33	401.10 ± 367.42	482.38 ± 486.44	161.92 ± 200.30	742.61 ± 918.40
NC_003923	0.33	366.62 ± 336.01	549.65 ± 544.99	142.38 ± 124.34	157.45 ± 208.84

Continues on next page.

Average $\left(fSpacing \times bps\right)$ continued...

Accession	**Global**	$\left(fSpacing \times bps\right)$			
Number	**G+C**	**CDS**	**NC**	**rRNA**	**tRNA**
NC_005791	0.33	348.11 ± 325.43	515.10 ± 430.13	162.76 ± 152.73	548.70 ± 677.34
NC_003997	0.35	318.36 ± 296.51	455.32 ± 436.86	128.78 ± 126.93	177.86 ± 239.43
NC_004722	0.35	319.75 ± 297.74	451.21 ± 444.26	124.05 ± 112.96	179.71 ± 287.21
NC_003909	0.36	327.02 ± 315.81	387.27 ± 391.10	129.17 ± 119.95	173.51 ± 200.01
NC_004368	0.36	302.61 ± 277.17	440.05 ± 403.44	147.52 ± 120.62	163.84 ± 189.18
NC_003212	0.37	284.49 ± 257.65	434.21 ± 372.43	135.63 ± 111.24	223.14 ± 260.16
NC_004350	0.37	275.87 ± 262.11	390.11 ± 386.20	137.76 ± 114.91	183.99 ± 207.89
NC_000907	0.38	247.39 ± 226.27	321.35 ± 313.24	167.60 ± 160.99	299.87 ± 346.96
NC_002940	0.38	249.17 ± 228.45	307.15 ± 293.10	168.79 ± 161.59	337.16 ± 475.17
NC_003210	0.38	275.03 ± 252.34	438.76 ± 393.62	136.39 ± 111.69	217.32 ± 278.64
NC_003869	0.38	291.88 ± 276.71	363.39 ± 390.36	145.88 ± 138.25	200.13 ± 270.22
NC_004668	0.38	267.83 ± 250.39	393.90 ± 403.21	155.72 ± 145.78	177.73 ± 258.28
NC_002737	0.39	241.32 ± 227.88	321.29 ± 304.45	144.45 ± 118.09	221.85 ± 396.86
NC_000912	0.40	228.10 ± 212.01	290.25 ± 272.82	174.99 ± 152.42	148.00 ± 133.49
NC_002620	0.40	226.77 ± 195.14	290.61 ± 256.07	160.25 ± 138.94	266.85 ± 333.03
NC_002689	0.40	240.37 ± 224.37	310.32 ± 288.95	148.79 ± 150.37	407.74 ± 534.76
NC_000922	0.41	226.84 ± 202.14	304.97 ± 273.47	192.93 ± 187.80	228.93 ± 278.18
NC_002179	0.41	226.79 ± 204.14	302.00 ± 259.13	163.22 ± 161.20	231.09 ± 235.89
NC_003901	0.41	198.37 ± 198.59	345.54 ± 352.01	161.33 ± 166.10	353.40 ± 457.54
NC_005043	0.41	226.90 ± 202.23	315.44 ± 277.54	188.64 ± 183.50	231.34 ± 278.27
NC_000918	0.43	212.42 ± 201.63	265.79 ± 247.11	143.37 ± 122.78	401.11 ± 456.66
NC_004663	0.43	206.44 ± 203.20	358.60 ± 365.81	153.74 ± 125.61	216.79 ± 265.49
NC_000964	0.44	204.37 ± 199.54	300.94 ± 271.41	122.47 ± 97.72	153.70 ± 157.25
NC_000853	0.46	181.23 ± 169.86	217.65 ± 206.52	131.17 ± 102.91	235.29 ± 313.62
NC_003143	0.48	175.54 ± 172.96	228.28 ± 222.93	159.14 ± 141.88	166.19 ± 182.03
NC_004088	0.48	174.85 ± 170.34	219.64 ± 218.50	168.10 ± 154.79	178.34 ± 223.08
NC_000917	0.49	186.36 ± 180.61	221.51 ± 205.59	171.23 ± 158.20	298.52 ± 248.80
NC_000916	0.50	170.08 ± 168.72	232.84 ± 220.17	142.28 ± 129.64	269.20 ± 275.42
NC_004431	0.50	168.53 ± 161.18	190.63 ± 199.26	147.44 ± 122.78	162.21 ± 140.93
NC_000913	0.51	167.44 ± 160.96	214.00 ± 215.60	149.37 ± 120.45	184.21 ± 209.09
NC_002695	0.51	168.69 ± 165.22	210.04 ± 229.29	148.55 ± 118.53	179.07 ± 199.17
NC_004741	0.51	165.73 ± 158.42	187.97 ± 187.38	142.46 ± 134.75	170.05 ± 215.68

Continues on next page.

Average $\left(fSpacing \times bps\right)$ continued...

Accession	**Global**	$\left(fSpacing \times bps\right)$			
Number	**G+C**	**CDS**	**NC**	**rRNA**	**tRNA**
NC_003197	0.52	166.41 ± 161.59	201.43 ± 199.17	151.24 ± 122.39	146.58 ± 130.37
NC_003198	0.52	166.24 ± 161.66	203.91 ± 199.36	138.91 ± 123.56	165.89 ± 218.74
NC_004556	0.52	182.18 ± 182.42	215.55 ± 222.21	135.78 ± 117.89	195.46 ± 217.71
NC_000919	0.53	172.13 ± 165.20	179.18 ± 166.86	138.07 ± 104.62	183.29 ± 220.88
NC_002488	0.53	180.53 ± 186.72	208.15 ± 220.19	139.56 ± 117.08	182.37 ± 180.64
NC_002935	0.53	169.14 ± 167.12	190.32 ± 178.63	165.71 ± 174.28	140.01 ± 164.35
NC_002932	0.57	180.97 ± 191.11	196.71 ± 190.41	153.01 ± 137.06	163.44 ± 177.52
NC_004307	0.60	187.40 ± 197.87	186.96 ± 198.15	148.26 ± 136.56	149.02 ± 193.24
NC_002939	0.61	170.47 ± 174.21	182.94 ± 180.81	147.97 ± 126.82	157.73 ± 172.26
NC_004369	0.63	182.05 ± 188.23	182.67 ± 183.95	137.50 ± 142.79	146.92 ± 136.01
NC_003919	0.65	206.92 ± 219.64	202.19 ± 217.64	136.22 ± 119.85	125.94 ± 120.28
NC_005085	0.65	203.55 ± 220.21	194.23 ± 198.62	144.57 ± 132.13	144.49 ± 139.86
NC_002696	0.67	203.38 ± 209.98	181.83 ± 192.16	141.98 ± 125.77	136.40 ± 150.71
NC_002927	0.68	214.89 ± 230.34	196.17 ± 204.85	130.26 ± 105.38	134.18 ± 122.07
NC_002928	0.68	214.26 ± 231.59	195.30 ± 203.26	134.99 ± 118.88	138.86 ± 131.00
NC_002929	0.68	216.88 ± 232.95	197.65 ± 196.94	129.09 ± 104.60	153.12 ± 144.81

Table A.8: This table presents the average $\left(fSpacing \times bps\right)$ values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figures 5.13 and B.11 on pages 77 and 155, respectively.

A.10 Average *bps* values under alternate search parameters - see caption.

Accession Number	Global G+C	CDS *bps*	NC *bps*	rRNA *bps*	tRNA *bps*
NC_002162	0.25	18.01 ± 14.44	19.55 ± 15.34	11.73 ± 8.07	15.18 ± 10.94
NC_002528	0.26	15.81 ± 12.25	20.26 ± 16.16	13.00 ± 8.47	14.53 ± 11.10
NC_001318	0.29	17.42 ± 16.38	19.78 ± 17.77	14.79 ± 10.35	17.24 ± 16.72
NC_003366	0.29	14.89 ± 11.55	20.03 ± 16.40	12.64 ± 8.38	14.48 ± 10.70
NC_004557	0.29	14.38 ± 11.05	17.10 ± 13.45	13.72 ± 9.49	14.60 ± 10.85
NC_000909	0.31	14.56 ± 11.38	18.48 ± 15.85	14.17 ± 9.78	14.18 ± 8.89
NC_003030	0.31	14.53 ± 11.21	17.67 ± 14.03	12.09 ± 7.46	14.91 ± 10.79
NC_003103	0.32	15.31 ± 11.73	16.33 ± 12.60	12.49 ± 7.70	12.76 ± 8.47
NC_003106	0.33	14.71 ± 11.46	16.70 ± 13.15	14.05 ± 9.83	13.05 ± 8.14
NC_003923	0.33	15.75 ± 12.39	17.57 ± 13.96	13.03 ± 8.58	11.74 ± 7.88
NC_005791	0.33	14.87 ± 11.21	18.21 ± 13.97	14.91 ± 10.73	14.27 ± 11.5
NC_003997	0.35	14.42 ± 11.08	15.16 ± 11.77	12.37 ± 7.85	12.64 ± 8.93
NC_004722	0.35	14.38 ± 11.07	15.29 ± 11.84	12.40 ± 8.00	12.27 ± 8.64
NC_003909	0.36	14.52 ± 11.27	14.24 ± 11.00	12.40 ± 7.92	12.58 ± 8.55
NC_004368	0.36	15.57 ± 12.08	15.78 ± 12.00	11.75 ± 7.66	10.86 ± 7.96
NC_003212	0.37	14.76 ± 11.39	14.83 ± 11.13	12.73 ± 8.67	14.50 ± 9.57
NC_004350	0.37	15.59 ± 12.25	15.59 ± 11.97	13.08 ± 9.01	13.56 ± 9.20
NC_000907	0.38	15.65 ± 12.12	16.00 ± 12.27	13.38 ± 9.88	14.06 ± 10.40
NC_002940	0.38	15.19 ± 11.95	15.59 ± 12.19	12.31 ± 8.51	13.62 ± 10.27
NC_003210	0.38	14.78 ± 11.49	15.08 ± 11.48	12.82 ± 8.79	14.75 ± 10.02
NC_003869	0.38	13.10 ± 9.75	14.17 ± 10.61	12.86 ± 8.83	13.67 ± 9.76
NC_004668	0.38	14.71 ± 11.34	14.66 ± 11.05	12.94 ± 8.83	13.06 ± 9.70
NC_002737	0.39	15.21 ± 11.65	15.05 ± 11.40	12.47 ± 8.20	12.79 ± 10.33
NC_000912	0.40	16.15 ± 12.51	15.79 ± 12.21	11.27 ± 7.13	14.60 ± 10.50
NC_002620	0.40	14.54 ± 11.43	14.45 ± 11.09	11.19 ± 6.97	16.40 ± 11.90
NC_002689	0.40	13.63 ± 9.92	15.79 ± 11.91	13.02 ± 8.72	15.70 ± 11.26

Continues on next page.

Average *bps* continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	*bps*	*bps*	*bps*	*bps*
NC_000922	0.41	14.57 ± 11.17	15.47 ± 11.78	11.36 ± 7.23	13.79 ± 10.27
NC_002179	0.41	14.38 ± 10.95	14.43 ± 10.81	14.39 ± 10.43	14.64 ± 10.46
NC_003901	0.41	13.86 ± 10.14	15.47 ± 12.19	13.82 ± 9.93	14.27 ± 10.40
NC_005043	0.41	14.57 ± 11.15	15.66 ± 12.07	11.33 ± 7.20	13.78 ± 10.19
NC_000918	0.43	12.74 ± 9.11	13.72 ± 10.33	14.28 ± 9.81	13.42 ± 8.75
NC_004663	0.43	14.28 ± 10.86	15.32 ± 11.95	14.87 ± 11.08	13.30 ± 9.32
NC_000964	0.44	13.84 ± 10.22	14.09 ± 10.36	12.54 ± 8.89	13.16 ± 9.49
NC_000853	0.46	12.92 ± 9.26	13.72 ± 10.11	13.05 ± 8.64	12.88 ± 9.13
NC_003143	0.48	15.53 ± 12.48	14.84 ± 11.27	14.18 ± 10.09	14.37 ± 11.05
NC_004088	0.48	15.43 ± 12.24	15.08 ± 11.88	13.20 ± 8.61	13.31 ± 10.50
NC_000917	0.49	13.29 ± 9.62	14.87 ± 11.10	16.65 ± 11.77	15.15 ± 10.25
NC_000916	0.50	14.34 ± 10.75	15.23 ± 11.69	11.03 ± 6.65	13.82 ± 9.10
NC_004431	0.50	15.38 ± 11.99	15.03 ± 11.58	12.84 ± 8.50	14.34 ± 11.70
NC_000913	0.51	15.47 ± 12.05	14.73 ± 11.06	12.85 ± 8.32	12.97 ± 8.86
NC_002695	0.51	15.36 ± 11.94	14.84 ± 11.24	12.80 ± 8.30	13.56 ± 9.64
NC_004741	0.51	15.38 ± 11.96	14.98 ± 11.31	13.57 ± 8.79	14.52 ± 9.80
NC_003197	0.52	15.26 ± 11.91	14.70 ± 11.18	13.18 ± 8.73	13.35 ± 9.63
NC_003198	0.52	15.19 ± 11.76	14.43 ± 10.78	13.59 ± 8.95	14.48 ± 10.43
NC_004556	0.52	19.13 ± 18.46	19.06 ± 18.42	11.73 ± 7.63	14.66 ± 12.24
NC_000919	0.53	14.88 ± 12.04	14.87 ± 11.77	11.27 ± 7.27	16.27 ± 12.23
NC_002488	0.53	17.07 ± 16.29	16.32 ± 15.53	11.63 ± 7.55	13.85 ± 9.90
NC_002935	0.53	16.96 ± 15.36	16.08 ± 13.51	16.27 ± 13.25	13.80 ± 9.54
NC_002932	0.57	16.65 ± 13.44	15.00 ± 11.58	13.04 ± 9.02	12.66 ± 8.22
NC_004307	0.60	18.09 ± 15.32	16.72 ± 14.08	13.04 ± 8.58	12.94 ± 9.44
NC_002939	0.61	15.59 ± 12.17	14.61 ± 11.23	12.57 ± 8.53	14.22 ± 10.63
NC_004369	0.63	17.01 ± 13.77	15.83 ± 12.52	14.42 ± 11.31	14.40 ± 9.68
NC_003919	0.65	19.58 ± 16.68	18.00 ± 14.92	14.45 ± 10.26	13.86 ± 9.75
NC_005085	0.65	18.69 ± 16.03	16.40 ± 13.62	13.16 ± 8.94	14.22 ± 9.30

Continues on next page.

Average *bps* continued...

Accession Number	**Global G+C**	**CDS** *bps*	**NC** *bps*	**rRNA** *bps*	**tRNA** *bps*
NC_002696	0.67	19.21 ± 15.88	16.47 ± 13.32	14.22 ± 10.27	14.13 ± 9.90
NC_002927	0.68	19.95 ± 17.03	16.48 ± 13.39	13.25 ± 9.44	13.56 ± 9.09
NC_002928	0.68	19.87 ± 16.95	16.35 ± 13.06	12.46 ± 8.58	13.35 ± 8.93
NC_002929	0.68	20.15 ± 17.21	16.67 ± 13.54	13.19 ± 9.42	13.58 ± 9.37

Table A.9: This table presents the average *bps* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The only exceptions are that the minimum GC base pair content is set to 0% and the maximum GU base pair content is set to 100%. The values in this table are graphed in Figure 7.1 on page 101.

A.11 Average *cSpacing* data under alternate search parameters - see caption.

Accession	Global	CDS	NC	rRNA	tRNA
Number	G+C	*cSpacing*	*cSpacing*	*cSpacing*	*cSpacing*
NC_002162	0.25	51.58 ± 23.35	53.63 ± 24.49	41.62 ± 14.79	46.77 ± 17.98
NC_002528	0.26	47.92 ± 20.06	54.14 ± 25.59	43.63 ± 14.79	45.50 ± 18.31
NC_001318	0.29	50.71 ± 25.98	54.33 ± 27.90	46.62 ± 17.55	49.07 ± 25.95
NC_003366	0.29	46.72 ± 19.15	54.57 ± 25.79	43.58 ± 14.92	45.44 ± 18.41
NC_004557	0.29	46.09 ± 18.45	50.16 ± 21.66	45.25 ± 16.49	45.99 ± 17.58
NC_000909	0.31	46.90 ± 18.90	52.56 ± 25.04	45.16 ± 16.75	45.39 ± 14.75
NC_003030	0.31	46.43 ± 18.71	51.18 ± 22.59	41.99 ± 13.52	46.64 ± 17.76
NC_003103	0.32	47.84 ± 19.56	49.18 ± 20.58	43.25 ± 14.08	42.38 ± 15.16
NC_003106	0.33	47.17 ± 19.19	49.73 ± 21.63	45.52 ± 16.86	43.27 ± 13.94
NC_003923	0.33	48.42 ± 20.46	51.06 ± 22.53	43.46 ± 15.06	41.14 ± 13.97
NC_005791	0.33	46.76 ± 18.71	51.67 ± 22.53	46.41 ± 17.75	45.32 ± 19.91
NC_003997	0.35	46.51 ± 18.68	47.36 ± 19.47	42.83 ± 14.39	42.36 ± 15.69
NC_004722	0.35	46.41 ± 18.64	47.55 ± 19.54	42.91 ± 14.63	42.05 ± 15.32
NC_003909	0.36	46.63 ± 18.89	46.12 ± 18.43	42.79 ± 14.50	42.27 ± 14.86
NC_004368	0.36	47.94 ± 19.98	48.04 ± 19.74	41.46 ± 13.85	39.65 ± 14.12
NC_003212	0.37	47.12 ± 19.07	46.71 ± 18.47	43.02 ± 15.09	45.22 ± 16.13
NC_004350	0.37	48.07 ± 20.21	47.72 ± 19.63	44.05 ± 15.86	42.91 ± 15.79
NC_000907	0.38	48.27 ± 20.07	48.85 ± 20.24	45.32 ± 16.97	44.74 ± 17.37
NC_002940	0.38	47.78 ± 19.83	48.07 ± 20.03	42.55 ± 14.85	45.06 ± 18.51
NC_003210	0.38	47.12 ± 19.18	47.17 ± 19.09	43.24 ± 15.31	45.47 ± 16.73
NC_003869	0.38	44.55 ± 16.74	45.82 ± 17.81	43.99 ± 16.02	43.43 ± 16.50
NC_004668	0.38	46.75 ± 19.00	46.31 ± 18.35	44.04 ± 16.29	43.17 ± 16.72
NC_002737	0.39	47.47 ± 19.38	47.03 ± 18.96	43.12 ± 14.77	42.70 ± 18.50
NC_000912	0.40	49.03 ± 20.59	48.08 ± 20.21	41.34 ± 13.54	45.38 ± 18.01
NC_002620	0.40	46.10 ± 18.93	45.45 ± 18.35	41.26 ± 13.15	48.69 ± 19.45
NC_002689	0.40	45.43 ± 17.11	48.51 ± 19.85	44.03 ± 15.41	47.28 ± 19.85

Continues on next page.

Average *cSpacing* continued...

Accession Number	**Global G+C**	**CDS** *cSpacing*	**NC** *cSpacing*	**rRNA** *cSpacing*	**tRNA** *cSpacing*
NC_000922	0.41	46.38 ± 18.64	47.48 ± 19.41	41.51 ± 13.26	44.61 ± 17.34
NC_002179	0.41	46.18 ± 18.33	46.01 ± 18.05	45.31 ± 17.85	45.75 ± 17.10
NC_003901	0.41	45.53 ± 17.48	48.56 ± 20.72	44.88 ± 17.02	45.37 ± 17.75
NC_005043	0.41	46.37 ± 18.62	47.81 ± 19.82	41.41 ± 13.20	44.47 ± 17.20
NC_000918	0.43	43.80 ± 15.86	45.70 ± 17.67	45.29 ± 16.75	44.23 ± 15.01
NC_004663	0.43	46.20 ± 18.32	47.84 ± 19.61	47.46 ± 18.73	43.24 ± 16.41
NC_000964	0.44	45.67 ± 17.47	45.81 ± 17.63	42.38 ± 15.74	44.03 ± 16.14
NC_000853	0.46	43.77 ± 16.02	45.13 ± 17.13	42.88 ± 14.48	43.40 ± 15.92
NC_003143	0.48	48.21 ± 20.64	47.22 ± 18.91	45.91 ± 17.15	46.69 ± 19.12
NC_004088	0.48	48.06 ± 20.26	47.59 ± 19.74	44.55 ± 14.90	44.14 ± 17.48
NC_000917	0.49	45.40 ± 16.75	47.48 ± 18.77	48.88 ± 20.15	47.10 ± 18.86
NC_000916	0.50	46.31 ± 18.24	47.79 ± 19.55	42.36 ± 13.92	44.46 ± 16.13
NC_004431	0.50	47.98 ± 19.89	47.36 ± 19.32	43.84 ± 14.95	46.32 ± 19.70
NC_000913	0.51	48.15 ± 19.98	46.81 ± 18.59	43.93 ± 14.74	44.13 ± 15.71
NC_002695	0.51	48.01 ± 19.86	47.09 ± 18.83	44.07 ± 14.71	45.39 ± 17.41
NC_004741	0.51	47.93 ± 19.86	47.39 ± 18.95	44.52 ± 15.24	45.51 ± 17.31
NC_003197	0.52	47.73 ± 19.78	46.80 ± 18.77	44.64 ± 15.48	44.04 ± 16.26
NC_003198	0.52	47.65 ± 19.58	46.37 ± 18.24	44.47 ± 15.52	45.70 ± 17.45
NC_004556	0.52	54.42 ± 29.21	54.79 ± 29.23	41.96 ± 13.80	47.24 ± 20.63
NC_000919	0.53	46.96 ± 19.80	47.22 ± 19.32	42.11 ± 14.16	49.11 ± 20.59
NC_002488	0.53	51.46 ± 26.02	50.82 ± 25.16	41.75 ± 13.74	46.21 ± 18.13
NC_002935	0.53	50.55 ± 24.68	49.38 ± 22.17	49.06 ± 21.56	44.93 ± 16.93
NC_002932	0.57	50.02 ± 22.08	47.58 ± 19.45	43.96 ± 15.84	42.85 ± 14.77
NC_004307	0.60	52.40 ± 24.82	50.38 ± 22.92	44.71 ± 15.49	42.90 ± 16.49
NC_002939	0.61	48.12 ± 20.22	46.77 ± 18.80	43.49 ± 15.47	46.25 ± 18.90
NC_004369	0.63	50.43 ± 22.44	48.63 ± 20.58	45.38 ± 18.10	46.38 ± 17.97
NC_003919	0.65	54.55 ± 26.72	52.24 ± 24.14	45.48 ± 17.21	45.43 ± 17.78
NC_005085	0.65	53.06 ± 25.71	49.63 ± 22.26	43.83 ± 15.30	45.89 ± 16.45

Continues on next page.

Average *cSpacing* continued...

Accession Number	**Global G+C**	**CDS** *cSpacing*	**NC** *cSpacing*	**rRNA** *cSpacing*	**tRNA** *cSpacing*
NC_002696	0.67	54.02 ± 25.61	49.57 ± 21.80	46.21 ± 17.95	45.50 ± 17.35
NC_002927	0.68	55.04 ± 27.20	49.55 ± 21.92	43.50 ± 16.51	45.02 ± 16.63
NC_002928	0.68	54.90 ± 27.10	49.35 ± 21.41	42.13 ± 15.33	44.76 ± 16.38
NC_002929	0.68	55.39 ± 27.47	49.99 ± 22.14	43.37 ± 16.47	45.02 ± 16.46

Table A.10: This table presents the average *cSpacing* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The only exceptions are that the minimum GC base pair content is set to 0% and the maximum GU base pair content is set to 100%. The values in this table are graphed in Figure 7.2 on page 102.

A.12 Average *fSpacing* under alternate search parameters - see caption.

Accession Number	Global G+C	CDS *fSpacing*	NC *fSpacing*	rRNA *fSpacing*	tRNA *fSpacing*
NC_002162	0.25	10.31 ± 5.04	10.28 ± 5.28	9.89 ± 4.91	9.51 ± 4.43
NC_002528	0.26	10.17 ± 5.09	9.88 ± 4.88	10.03 ± 4.95	10.02 ± 5.36
NC_001318	0.29	10.47 ± 5.69	10.66 ± 5.74	9.65 ± 4.58	9.41 ± 5.26
NC_003366	0.29	10.42 ± 5.20	10.43 ± 5.28	10.13 ± 4.57	9.29 ± 4.66
NC_004557	0.29	10.57 ± 5.41	10.54 ± 5.43	10.19 ± 4.76	10.07 ± 4.66
NC_000909	0.31	11.15 ± 6.02	10.77 ± 5.93	10.01 ± 5.22	9.67 ± 4.72
NC_003030	0.31	10.54 ± 5.39	10.57 ± 5.47	10.31 ± 4.98	9.93 ± 4.89
NC_003103	0.32	10.43 ± 5.33	10.44 ± 5.34	10.32 ± 5.62	9.78 ± 4.62
NC_003106	0.33	10.77 ± 5.78	10.41 ± 5.60	10.36 ± 5.15	9.63 ± 4.69
NC_003923	0.33	10.43 ± 5.23	10.42 ± 5.24	9.75 ± 4.93	11.75 ± 4.19
NC_005791	0.33	10.30 ± 5.05	10.24 ± 5.34	9.98 ± 5.25	9.66 ± 5.03
NC_003997	0.35	10.63 ± 5.67	10.69 ± 5.88	10.21 ± 4.72	8.98 ± 4.25
NC_004722	0.35	10.60 ± 5.71	10.64 ± 5.79	10.21 ± 4.67	9.53 ± 4.65
NC_003909	0.36	10.64 ± 5.85	10.69 ± 6.00	10.17 ± 4.70	9.37 ± 4.46
NC_004368	0.36	10.30 ± 5.24	10.35 ± 5.29	10.31 ± 4.98	9.30 ± 4.30
NC_003212	0.37	10.65 ± 5.64	10.57 ± 5.66	9.95 ± 4.69	9.45 ± 4.50
NC_004350	0.37	10.35 ± 5.34	10.30 ± 5.41	9.97 ± 5.22	9.40 ± 4.59
NC_000907	0.38	10.39 ± 5.37	10.49 ± 5.70	11.13 ± 6.46	9.98 ± 4.97
NC_002940	0.38	10.57 ± 5.42	10.40 ± 5.43	10.13 ± 5.22	10.77 ± 5.67
NC_003210	0.38	10.62 ± 5.66	10.66 ± 5.78	10.01 ± 4.70	9.27 ± 4.30
NC_003869	0.38	10.98 ± 6.11	10.83 ± 5.94	10.61 ± 5.52	9.06 ± 4.57
NC_004668	0.38	10.47 ± 5.41	10.54 ± 5.74	10.54 ± 6.11	9.52 ± 4.58
NC_002737	0.39	10.35 ± 5.25	10.32 ± 5.39	10.37 ± 5.14	9.39 ± 4.38
NC_000912	0.40	10.47 ± 5.21	10.49 ± 5.42	10.80 ± 5.45	8.95 ± 4.15
NC_002620	0.40	10.34 ± 5.59	10.20 ± 5.48	10.47 ± 5.04	9.99 ± 4.75
NC_002689	0.40	10.54 ± 5.50	10.35 ± 5.39	10.50 ± 4.99	9.94 ± 4.84

Continues on next page.

Average *fSpacing* continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	*fSpacing*	*fSpacing*	*fSpacing*	*fSpacing*
NC_000922	0.41	10.41 ± 5.55	10.29 ± 5.61	10.47 ± 5.07	9.73 ± 5.22
NC_002179	0.41	10.45 ± 5.66	10.39 ± 5.62	9.42 ± 4.48	9.36 ± 4.78
NC_003901	0.41	10.42 ± 5.80	11.01 ± 7.16	10.01 ± 4.96	10.27 ± 5.39
NC_005043	0.41	10.39 ± 5.53	10.38 ± 5.73	10.41 ± 5.06	9.71 ± 5.22
NC_000918	0.43	10.84 ± 5.97	11.22 ± 6.50	9.52 ± 4.30	10.15 ± 5.59
NC_004663	0.43	10.39 ± 5.46	10.74 ± 5.89	10.30 ± 4.90	9.39 ± 4.83
NC_000964	0.44	10.53 ± 5.57	10.68 ± 5.93	9.78 ± 4.64	9.69 ± 4.59
NC_000853	0.46	10.52 ± 5.62	10.59 ± 5.87	9.85 ± 4.56	10.46 ± 6.11
NC_003143	0.48	10.45 ± 5.38	10.70 ± 5.77	10.26 ± 5.42	10.59 ± 5.61
NC_004088	0.48	10.47 ± 5.40	10.67 ± 5.72	10.51 ± 5.54	10.53 ± 5.38
NC_000917	0.49	11.21 ± 6.30	10.95 ± 6.11	10.25 ± 4.10	10.49 ± 3.91
NC_000916	0.50	10.41 ± 5.45	10.54 ± 5.86	11.90 ± 7.98	9.91 ± 5.14
NC_004431	0.50	10.42 ± 5.33	10.45 ± 5.42	10.34 ± 5.38	10.37 ± 5.29
NC_000913	0.51	10.43 ± 5.34	10.46 ± 5.53	10.41 ± 5.38	10.24 ± 5.32
NC_002695	0.51	10.45 ± 5.38	10.55 ± 5.56	10.67 ± 5.55	10.78 ± 5.95
NC_004741	0.51	10.37 ± 5.28	10.49 ± 5.44	9.58 ± 4.89	9.80 ± 5.16
NC_003197	0.52	10.34 ± 5.24	10.44 ± 5.51	10.71 ± 5.75	9.84 ± 4.88
NC_003198	0.52	10.35 ± 5.27	10.46 ± 5.53	9.80 ± 4.88	9.93 ± 5.09
NC_004556	0.52	10.99 ± 6.22	11.51 ± 7.41	10.48 ± 5.24	11.06 ± 6.63
NC_000919	0.53	10.58 ± 5.83	11.00 ± 6.42	11.28 ± 6.18	10.99 ± 6.70
NC_002488	0.53	11.24 ± 6.47	11.75 ± 7.82	10.52 ± 5.30	11.10 ± 5.95
NC_002935	0.53	10.57 ± 5.61	10.91 ± 6.10	10.45 ± 5.52	10.14 ± 5.29
NC_002932	0.57	10.35 ± 5.30	10.78 ± 6.04	10.25 ± 5.13	9.85 ± 5.03
NC_004307	0.60	10.45 ± 5.28	10.84 ± 5.79	10.96 ± 5.50	9.52 ± 4.89
NC_002939	0.61	10.23 ± 5.18	10.58 ± 5.63	10.35 ± 4.98	10.37 ± 4.96
NC_004369	0.63	10.35 ± 5.21	10.59 ± 5.64	9.90 ± 4.75	10.24 ± 5.56
NC_003919	0.65	10.41 ± 5.15	10.57 ± 5.43	9.80 ± 4.89	9.80 ± 4.92
NC_005085	0.65	10.40 ± 5.18	10.65 ± 5.70	9.99 ± 4.86	10.20 ± 4.87

Continues on next page.

Average *fSpacing* continued...

Accession Number	**Global G+C**	**CDS** *fSpacing*	**NC** *fSpacing*	**rRNA** *fSpacing*	**tRNA** *fSpacing*
NC_002696	0.67	10.30 ± 5.06	10.41 ± 5.33	10.14 ± 4.48	9.89 ± 4.78
NC_002927	0.68	10.37 ± 5.12	10.49 ± 5.40	9.87 ± 4.79	10.17 ± 5.15
NC_002928	0.68	10.36 ± 5.13	10.49 ± 5.39	10.12 ± 5.00	10.49 ± 5.45
NC_002929	0.68	10.40 ± 5.14	10.55 ± 5.50	9.82 ± 4.77	9.83 ± 5.05

Table A.11: This table presents the average *fSpacing* values and their standard deviations measured in the various genomic domains of the genomes in our test set. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The only exceptions are that the minimum GC base pair content is set to 0% and the maximum GU base pair content is set to 100%. The values in this table are graphed in Figure 7.3 on page 103.

Appendix B

Stem-loop Metric Results on Shuffled Genomes

This appendix presents the results obtained when the various stem-loop metrics were measured on random sequences. The random sequences were generated by shuffling each of the 58 genomes in our working set. To shuffle a genome, a nucleotide within the sequence is randomly selected. This nucleotide is appended onto the end of another string (call it `shuffled_string`, for instance). This is repeated until each nucleotide has been randomly purged from the genomic sequence and appended onto the `shuffled_string`.

The graphs presented in Chapter 5 were generated by plotting four points for each of the genomes. One for each of the genomic domains - CDS, NC, rRNA, and tRNA. Now in addition to the CDS, NC, rRNA, and tRNA points for each genome, there is a 5^{th} data point. It depicts the results obtained after randomly shuffling the genome.

B.1 Base Pairs Metric - *bps*

The plot for the average *bps* in randomly shuffled sequences is included in Figure B.1. The random values appear to roughly correspond with the NC plot more than the CDS, rRNA, or tRNA plots. The random *bps* plot intersects with the rRNA plot at roughly 52% *global* G+C content. This is the most significant observation in Figure B.1.

Recall, the genomic sequence and its shuffled or randomized counterpart have the same *global* G+C content. In Chapter 5, Figure 5.1 showed that the G+C content levels are not

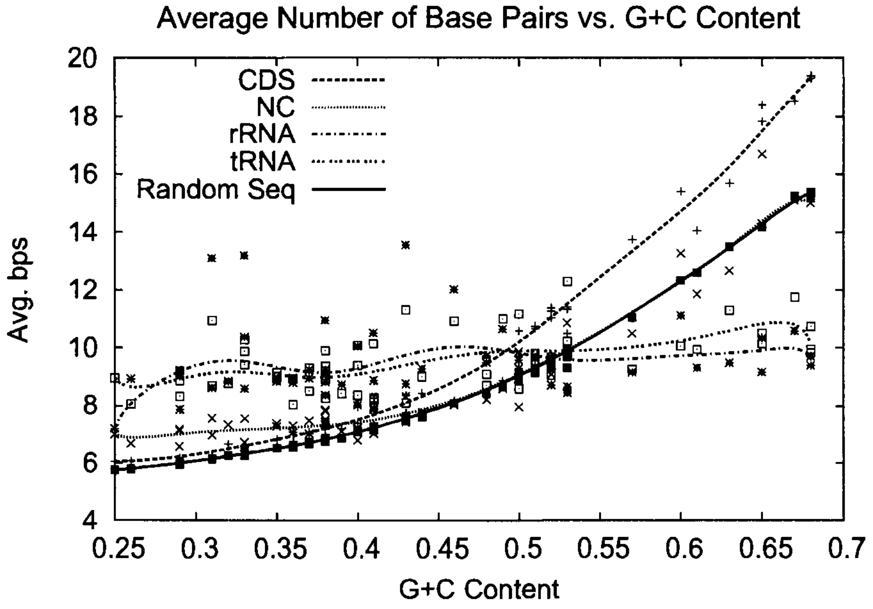


Figure B.1: Average number *bps* in stems-loops arising from various genomic domains. In addition to the plots in Figure 5.2 this graph shows the average *bps* found in random sequences across the G+C content spectrum. The corresponding data is presented in Appendices A.4 and C.1.

uniform across the genomic domains - CDS, NC, rRNA, and tRNA - within most genomic sequences. This is especially true of A+T rich and G+C rich sequences. The CDS domains comprise the vast majority of the bacterial genomes (roughly 90% or more). Therefore, the G+C content in the randomized genomic sequences will more closely resemble CDS. The differences observed in metric values between the genomic domains and the randomized genomic sequences is in large part related to differences in G+C content levels. This caveat is important to remember when studying the following figures.

The average G+C content in structural RNA genes is approximately 53% (Table 5.1). When the G+C content in random sequences and the genomic domains is roughly equivalent, the difference in their respective *bps* values diminishes greatly (Figure B.1). Interestingly, at 50-55% G+C, the bps_{CDS} values deviate more from bps_{random} then bps_{rRNA}. This suggests that the *bps* metric is more apt at distinguishing CDS regions from random sequences than it is at distinguishing structural RNAs from random sequences when the G+C content is 50-55%.

A more rigorous statistical comparison between random sequences and genomic sequences can be based on the Normal Standard Distribution. This involves calculating the Z Score for each genomic domain. The Z Score indicates how many standard deviations the mean (e.g. $\overline{x}_{CDS}$) deviates from the random mean, $\overline{x}_{random}$. Suppose $Z_{NC} = 1.0$, this indicates its average value for the NC domain is 1 standard deviation greater than the random sequence mean. Likewise, suppose $Z_{rRNA} = -2.0$, this indicates the average rRNA value is 2 standard deviations less than the random sequence average. Generally, a "significant" Z Score occurs when $Z \leq -3.0$ or $Z \geq 3.0$.

Figure B.2 depicts the Z Scores for the genomic domains with respect to random sequences. As depicted in the earlier figure, the G+C content is equivalent, there is less disparity between the respective *bps* values with the exception of CDS (Figure B.2 and Appendix C.7). In most cases when the *global* G+C content is 50-55%, the Z_{rRNA} values are significant (Appendix C.7). However, the Z_{random} and Z_{CDS} or Z_{NC} are typically more significant.

B.2 Span Metric - *span*

The results attained using the *span* metric on the randomly shuffled genomes are shown in Figure B.3. The Normal Standardized Z Scores are depicted in Figure B.4. These graphs

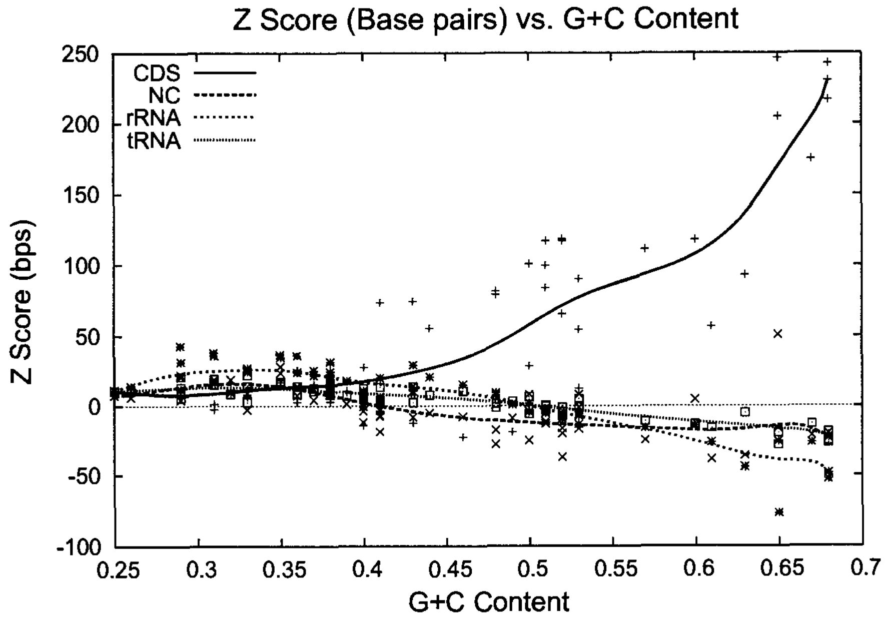


Figure B.2: *bps Z* Scores relative to random sequences. Random sequences were generated by randomly shuffling the genomic sequences in our working set. The corresponding data is presented in Appendix C.7.

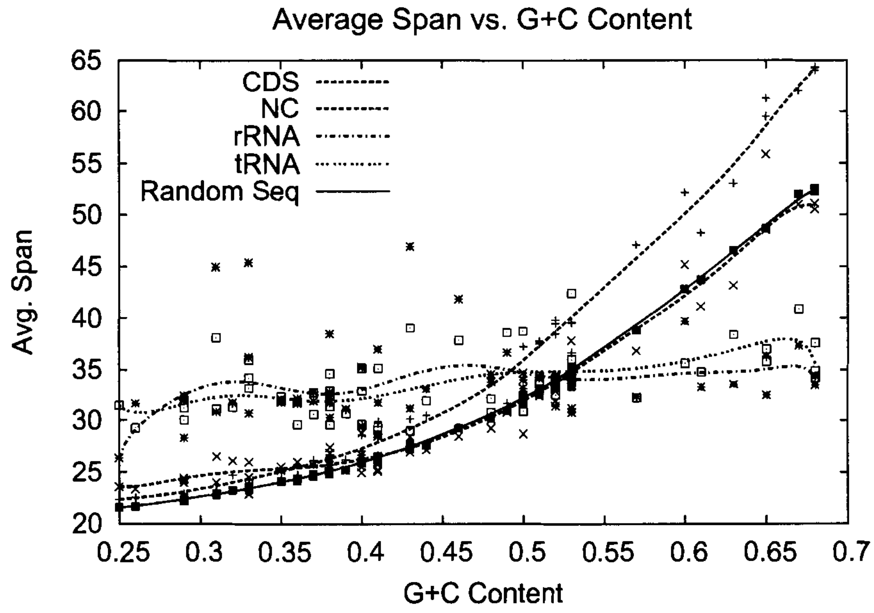


Figure B.3: Average Stem-loop *span* vs. G+C content including random sequences. The corresponding data is presented in Appendices A.5 and C.2.

and their trends mirror those of the *bps* metric. Rather than repeating the explanation let's look at the next metric.

B.3 Center Point Spacing Metric - *cSpacing*

The average *cSpacing* values observed in the randomly shuffled sequences are depicted in Figure B.5. These values appear to loosely coincide with those for the CDS and NC domains. The rRNA *cSpacing* values are lower than the random sequence *cSpacing* values with only one exception - NC_002935, which has a *global* G+C content of 53% (Appendix A.6). Figure B.6 conveys the same information in terms of Normal Standardized Z Scores. Note, the $cSpacing_{rRNA}$ values continue to differ significantly from the random sequences when their G+C content levels are similar (Appendix C.9). Interestingly, at 50-55% G+C, the Z_{CDS} and Z_{NC} values are typically more significant than the Z_{rRNA} values.

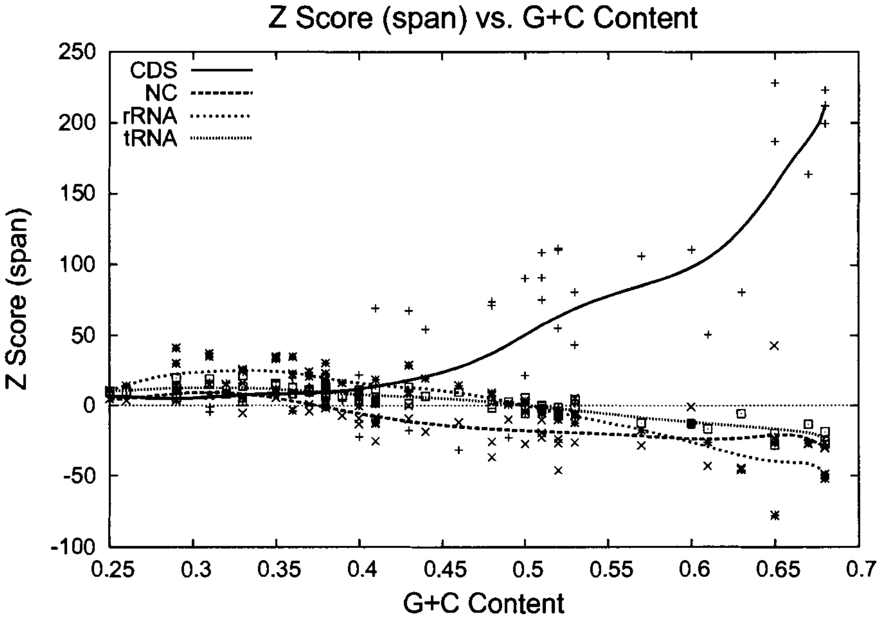


Figure B.4: *span* *Z* Scores relative to random sequences. The corresponding data is presented in Appendix C.8.

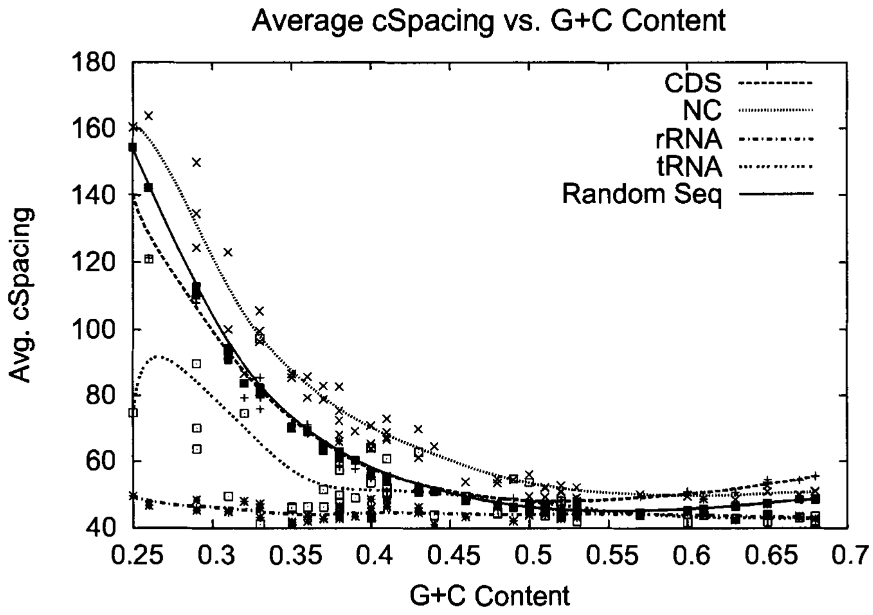


Figure B.5: Average stem-loop *cSpacing* vs. G+C content - including random sequences. The corresponding data is presented in Appendices A.6 and C.3

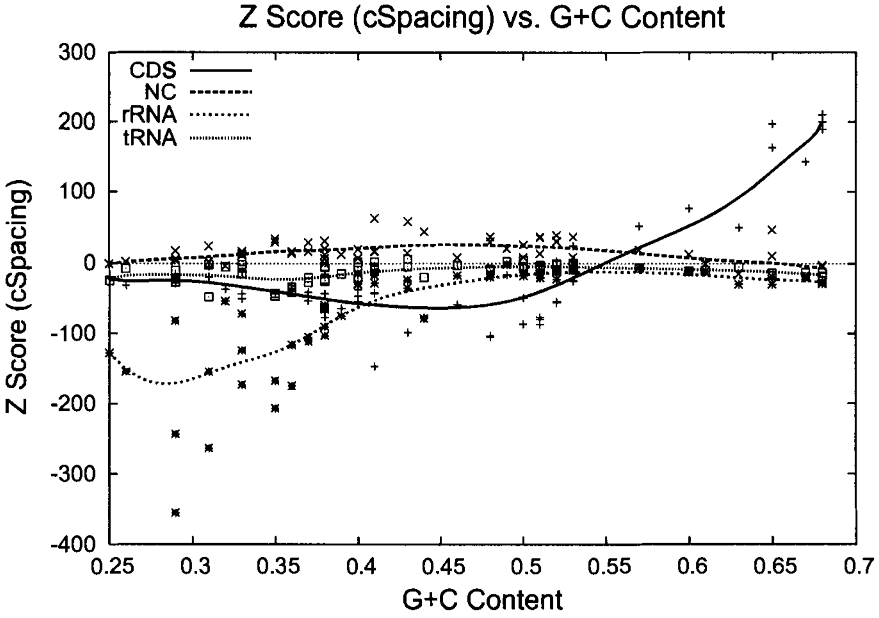


Figure B.6: *cSpacing Z* Scores relative to random sequences. The corresponding data is presented in Appendix C.9.

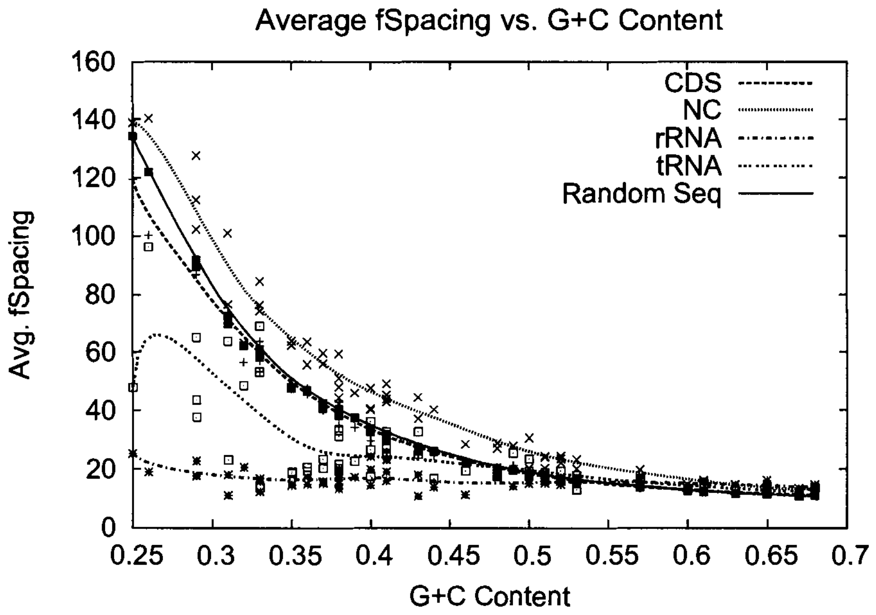


Figure B.7: Average *fSpacing* vs. G+C content - including random sequences. The corresponding data is presented in Appendices A.7 and C.4.

B.4 Foot Spacing Metric - *fSpacing*

The average *fSpacing* values for randomly shuffled sequences are plotted in Figure B.7. The Normal Standardized Z Scores for the *fSpacing* metric are plotted in Figure B.8. When the G+C content is 50-55%, the $fSpacing_{rRNA}$ values are usually significant (Appendix C.10). At the same G+C content levels, the Z Scores for the CDS and NC domains are typically more significant.

B.5 $(cSpacing \times bps)$ Metric

The average $\left(cSpacing \times bps\right)$ values in random sequences are shown in Figure B.9. The Normalized Z Scores are plotted in Figure B.10. The Z_{rRNA} values are significantly negative when the difference in G+C content between rRNA and the random sequence is negligible (Appendix C.11). Conversely, the Z_{CDS} values are significantly positive under the same

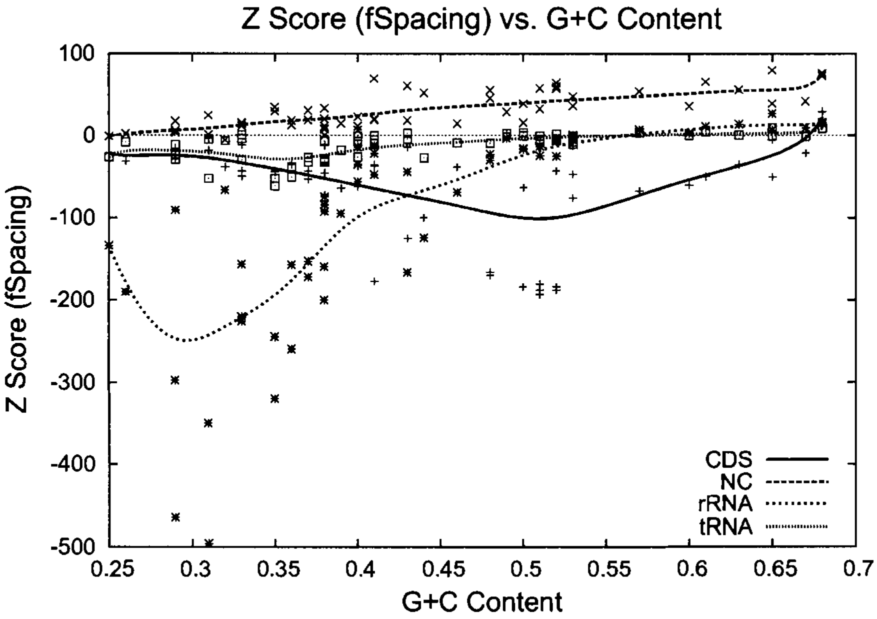


Figure B.8: *fSpacing Z* Scores relative to random sequences. The corresponding data is presented in Appendix C.10.

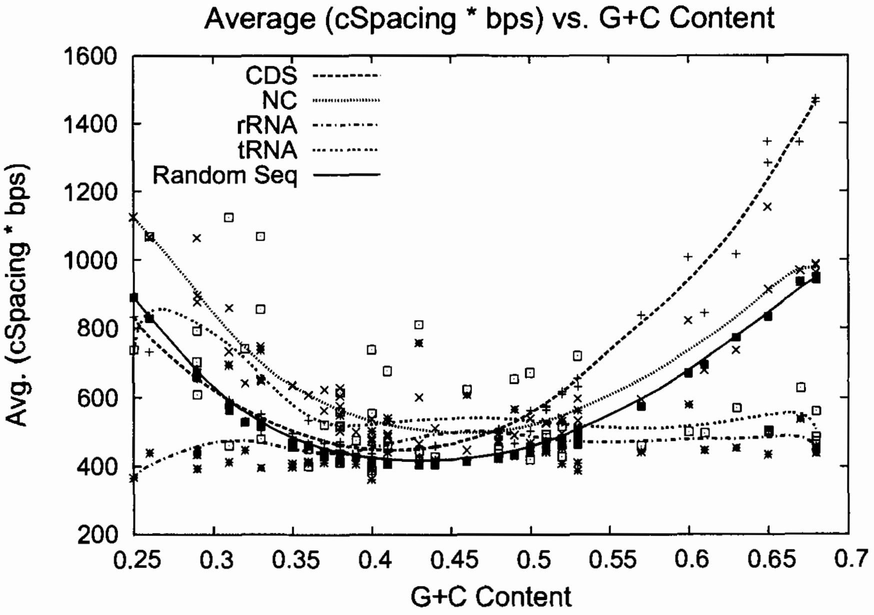


Figure B.9: Average $\left(cSpacing \times bps\right)$ vs. G+C content - including random sequences

circumstances.

B.6 $(fSpacing \times bps)$ Metric

The average $\left(fSpacing \times bps\right)$ values observed in randomly shuffled sequences over the G+C content spectrum is depicted in Figure B.11. The Normalized Z Scores are plotted in Figure B.10. The results are similar to the previous metric. The Z_{rRNA} values are significantly negative when the difference in G+C content between rRNA and the random sequence is negligible (Appendix C.11). Conversely, the Z_{CDS} values are significantly positive under the same circumstances.

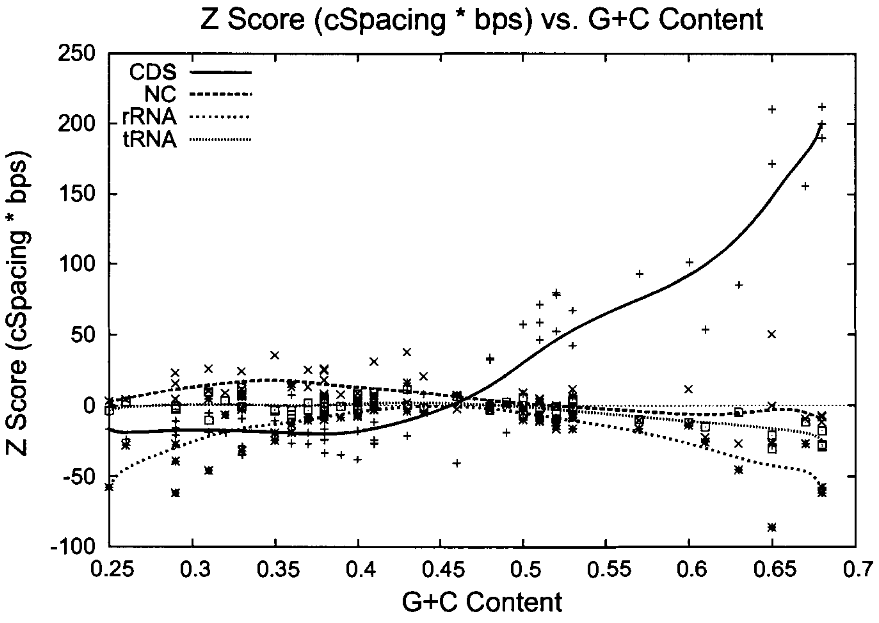


Figure B.10: $\left(cSpacing \times bps\right)$ Z Scores relative to random sequences. The corresponding data is presented in Appendix C.11.

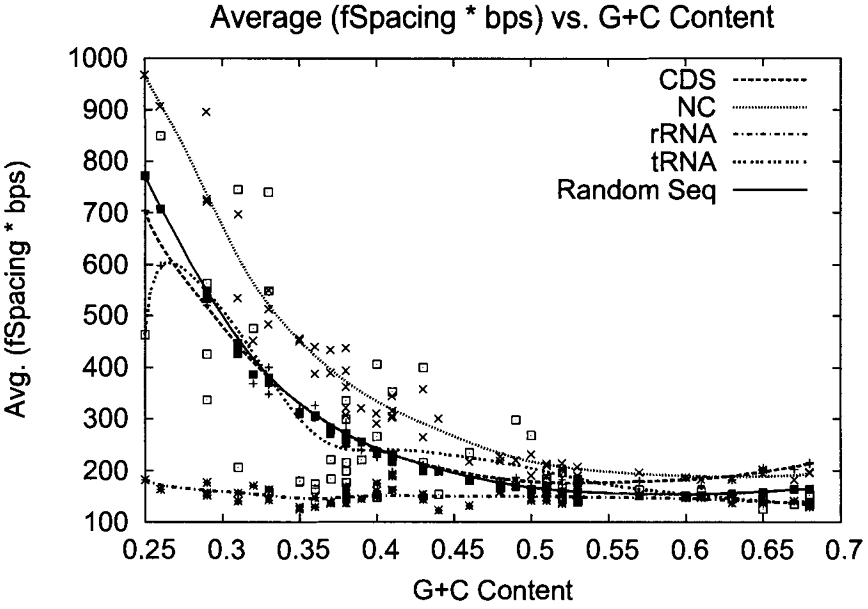


Figure B.11: Average $\left(fSpacing \times bps\right)$ vs. G+C content - including random sequences.

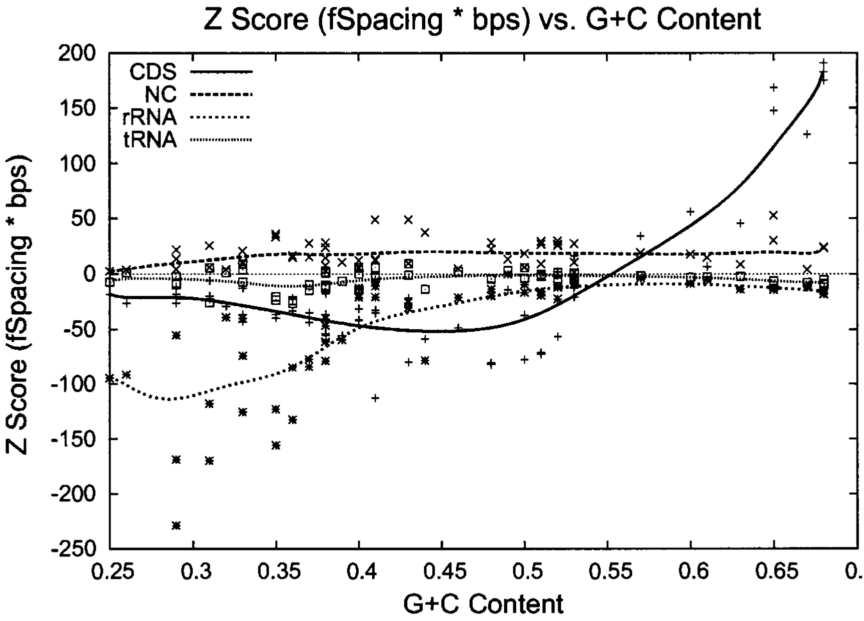


Figure B.12: $\left(fSpacing \times bps\right)$ Z Scores relative to random sequences. The corresponding data is presented in Appendix C.12.

B.7 Review

Z Scores for various metrics have been tabulated to study stem-loop metrics as they occur in random sequences. The point of interest occurs where the base composition is equivalent between naturally occurring sequences and their randomized counterparts. Here we know that differences seen in their metric values cannot be attributed to dicrepancies in base composition.

Typically, metric values found in rRNAs differ significantly from random sequences with the same base composition. Interestingly, the CDS and NC values commonly deviated more significantly from the random sequences than the rRNAs. This could indicate an importance in secondary structure along protein-encoding genes or transcripts. More study is required to understand the meaning or significance of these observations.

Appendix C

Stem-loop Metric Data on Shuffled Genomes

C.1 Average *bps* in Random Sequences Across G+C Content Spectrum

GlobalG+C	Average *bps*
0.25	5.75 ± 1.72
0.26	5.79 ± 1.78
0.29	5.93 ± 1.93
0.29	5.95 ± 1.95
0.29	5.96 ± 1.92
0.31	6.11 ± 2.11
0.31	6.14 ± 2.16
0.32	6.25 ± 2.24
0.33	6.25 ± 2.26
0.33	6.26 ± 2.26
0.33	6.32 ± 2.31
0.35	6.51 ± 2.52
0.35	6.52 ± 2.53

Continues on next page.

Average *bps* continued...

Global G+C	**Average *bps***
0.36	6.53 ± 2.54
0.36	6.55 ± 2.54
0.37	6.66 ± 2.68
0.37	6.74 ± 2.78
0.38	6.75 ± 2.78
0.38	6.77 ± 2.80
0.38	6.79 ± 2.84
0.38	6.81 ± 2.85
0.38	6.83 ± 2.88
0.39	6.87 ± 2.93
0.40	7.08 ± 3.15
0.40	7.08 ± 3.10
0.40	7.13 ± 3.22
0.41	7.16 ± 3.21
0.41	7.16 ± 3.22
0.41	7.17 ± 3.25
0.41	7.28 ± 3.35
0.43	7.49 ± 3.58
0.43	7.61 ± 3.70
0.44	7.61 ± 3.70
0.46	8.10 ± 4.24
0.48	8.42 ± 4.56
0.48	8.42 ± 4.57
0.49	8.65 ± 4.82
0.50	8.85 ± 5.01
0.50	9.13 ± 5.33
0.51	9.14 ± 5.33
0.51	9.21 ± 5.40
0.51	9.27 ± 5.46

Continues on next page.

Average *bps* continued...

Global G+C	**Average *bps***
0.52	9.43 ± 5.62
0.52	9.56 ± 5.76
0.52	9.62 ± 5.83
0.53	9.31 ± 5.52
0.53	9.70 ± 5.93
0.53	9.98 ± 6.22
0.57	11.07 ± 7.31
0.60	12.35 ± 8.57
0.61	12.62 ± 8.82
0.63	13.51 ± 9.69
0.65	14.18 ± 10.30
0.65	14.20 ± 10.32
0.67	15.24 ± 11.30
0.68	15.32 ± 11.38
0.68	15.32 ± 11.39
0.68	15.39 ± 11.44

Table C.1: This table presents the average *bps* values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.1 on page 144.

C.2 Average *span* in Random Sequences Across G+C Content Spectrum

GlobalG+C	Average *span*
0.25	21.51 ± 6.25
0.26	21.60 ± 6.36
0.29	22.13 ± 6.78
0.29	22.19 ± 6.81
0.29	22.19 ± 6.89
0.31	22.72 ± 7.29
0.31	22.82 ± 7.45
0.32	23.20 ± 7.71
0.33	23.18 ± 7.73
0.33	23.22 ± 7.73
0.33	23.37 ± 7.88
0.35	24.03 ± 8.47
0.35	24.06 ± 8.50
0.36	24.09 ± 8.55
0.36	24.15 ± 8.54
0.37	24.56 ± 8.97
0.37	24.76 ± 9.22
0.38	24.81 ± 9.21
0.38	24.87 ± 9.28
0.38	24.96 ± 9.41
0.38	25.05 ± 9.52
0.38	25.00 ± 9.45
0.39	25.19 ± 9.66
0.40	25.86 ± 10.29
0.40	25.89 ± 10.18
0.40	26.03 ± 10.54
0.41	26.13 ± 10.51

Continues on next page.

Average *span* continued...

Global G+C	**Average** *span*
0.41	26.14 ± 10.62
0.41	26.16 ± 10.56
0.41	26.54 ± 10.96
0.43	27.20 ± 11.64
0.43	27.57 ± 11.96
0.44	27.57 ± 12.00
0.46	29.20 ± 13.66
0.48	30.21 ± 14.65
0.48	30.22 ± 14.70
0.49	30.95 ± 15.48
0.50	31.60 ± 16.10
0.50	32.51 ± 17.05
0.51	32.53 ± 17.05
0.51	32.75 ± 17.30
0.51	32.95 ± 17.46
0.52	33.48 ± 18.04
0.52	33.87 ± 18.46
0.52	34.08 ± 18.63
0.53	33.32 ± 17.85
0.53	34.39 ± 19.02
0.53	35.24 ± 19.92
0.57	38.72 ± 23.33
0.60	42.81 ± 27.40
0.61	43.72 ± 28.26
0.63	46.55 ± 31.02
0.65	48.72 ± 33.02
0.65	48.70 ± 33.00
0.67	51.98 ± 36.10
0.68	52.25 ± 36.33

Continues on next page.

Average *span* continued...

Global G+C	**Average** *span*
0.68	52.25 ± 36.35
0.68	52.49 ± 36.54

Table C.2: This table presents the average *span* values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.3 on page 147.

C.3 Average *cSpacing* in Random Sequences Across G+C Content Spectrum

GlobalG+C	**Average** *cSpacing*
0.25	154.50 ± 95.30
0.26	142.23 ± 83.77
0.29	110.28 ± 62.92
0.29	112.46 ± 65.01
0.29	112.54 ± 63.47
0.31	90.69 ± 49.39
0.31	93.22 ± 50.82
0.32	83.47 ± 43.22
0.33	80.31 ± 41.04
0.33	81.50 ± 41.68
0.33	82.22 ± 42.79
0.35	69.72 ± 33.49
0.35	70.31 ± 33.87
0.36	68.74 ± 32.80
0.36	69.19 ± 33.17
0.37	63.02 ± 28.69
0.37	64.96 ± 29.68

Continues on next page.

Average *cSpacing* continued...

Global G+C	**Average *cSpacing***
0.38	60.85 ± 26.64
0.38	61.25 ± 26.89
0.38	61.42 ± 27.44
0.38	62.88 ± 28.21
0.38	62.80 ± 28.18
0.39	60.28 ± 26.36
0.40	56.07 ± 23.44
0.40	56.70 ± 23.74
0.40	56.91 ± 24.26
0.41	53.68 ± 21.24
0.41	55.18 ± 22.31
0.41	55.29 ± 22.45
0.41	55.37 ± 22.54
0.43	50.55 ± 18.83
0.43	51.50 ± 19.61
0.44	50.51 ± 18.78
0.46	47.74 ± 16.67
0.48	46.45 ± 15.61
0.48	46.46 ± 15.68
0.49	45.87 ± 15.05
0.50	45.04 ± 14.49
0.50	45.28 ± 14.71
0.51	44.81 ± 14.25
0.51	44.81 ± 14.38
0.51	44.88 ± 14.35
0.52	44.44 ± 14.11
0.52	44.46 ± 14.10
0.52	44.50 ± 14.15
0.53	44.18 ± 14.00

Continues on next page.

Average *cSpacing* continued...

Global G+C	**Average** *cSpacing*
0.53	44.29 ± 14.05
0.53	44.42 ± 14.06
0.57	44.27 ± 14.43
0.60	45.04 ± 15.58
0.61	45.27 ± 15.87
0.63	46.15 ± 16.94
0.65	46.93 ± 17.74
0.65	46.90 ± 17.71
0.67	48.23 ± 19.06
0.68	48.22 ± 19.18
0.68	48.29 ± 19.17
0.68	48.36 ± 19.26

Table C.3: This table presents the average *cSpacing* values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.5 on page 149.

C.4 Average *fSpacing* in Random Sequences Across G+C Content Spectrum

GlobalG+C	Average *fSpacing*
0.25	134.27 ± 95.18
0.26	121.96 ± 83.68
0.29	89.55 ± 62.78
0.29	91.77 ± 64.90
0.29	91.80 ± 63.39
0.31	69.47 ± 49.16
0.31	72.08 ± 50.66
0.32	61.98 ± 43.03
0.33	58.69 ± 40.79
0.33	60.01 ± 41.43
0.33	60.75 ± 42.51
0.35	47.62 ± 33.18
0.35	48.21 ± 33.54
0.36	46.54 ± 32.46
0.36	47.05 ± 32.81
0.37	40.43 ± 28.19
0.37	42.50 ± 29.26
0.38	38.07 ± 26.06
0.38	38.53 ± 26.36
0.38	38.65 ± 26.92
0.38	40.15 ± 27.70
0.38	40.20 ± 27.67
0.39	37.40 ± 25.78
0.40	32.60 ± 22.65
0.40	33.40 ± 23.00
0.40	33.54 ± 23.58
0.41	29.91 ± 20.36

Continues on next page.

Average *fSpacing* continued...

Global G+C	**Average *fSpacing***
0.41	31.65 ± 21.59
0.41	31.78 ± 21.66
0.41	31.82 ± 21.78
0.43	26.12 ± 17.66
0.43	27.31 ± 18.51
0.44	26.10 ± 17.58
0.46	22.31 ± 15.01
0.48	20.44 ± 13.57
0.48	20.45 ± 13.49
0.49	19.42 ± 12.63
0.50	17.71 ± 11.38
0.50	18.47 ± 11.93
0.51	17.23 ± 10.90
0.51	17.35 ± 11.10
0.51	17.53 ± 11.17
0.52	16.25 ± 10.22
0.52	16.37 ± 10.28
0.52	16.65 ± 10.52
0.53	15.35 ± 9.43
0.53	15.94 ± 9.95
0.53	16.64 ± 10.50
0.57	13.58 ± 7.98
0.60	12.30 ± 6.87
0.61	12.04 ± 6.66
0.63	11.46 ± 6.14
0.65	11.13 ± 5.82
0.65	11.13 ± 5.85
0.67	10.75 ± 5.50
0.68	10.67 ± 5.43

Continues on next page.

Average *fSpacing* continued...

Global G+C	**Average *fSpacing***
0.68	10.69 ± 5.44
0.68	10.70 ± 5.44

Table C.4: This table presents the average *fSpacing* values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.7 on page 151.

C.5 Average $(cSpacing \times bps)$ in Random Sequences Across G+C Content Spectrum

GlobalG+C	**Average** $\left(cSpacing \times bps\right)$
0.25	893.11 ± 650.73
0.26	828.77 ± 582.24
0.29	661.45 ± 458.72
0.29	670.98 ± 462.76
0.29	674.69 ± 458.24
0.31	564.11 ± 394.53
0.31	575.32 ± 393.43
0.32	528.48 ± 358.53
0.33	515.87 ± 354.42
0.33	518.04 ± 350.99
0.33	521.83 ± 363.28
0.35	463.06 ± 315.75
0.35	466.57 ± 318.48
0.36	459.30 ± 313.53
0.36	461.42 ± 317.91
0.37	436.14 ± 310.08
0.37	443.01 ± 302.76

Continues on next page.

Average $\left(cSpacing \times bps\right)$ continued...

Global G+C	**Average** $\left(cSpacing \times bps\right)$
0.38	427.49 ± 298.04
0.38	427.67 ± 300.55
0.38	430.52 ± 303.31
0.38	435.23 ± 303.60
0.38	437.73 ± 306.89
0.39	427.07 ± 303.62
0.40	416.11 ± 310.10
0.40	416.50 ± 302.55
0.40	417.68 ± 302.23
0.41	407.46 ± 301.27
0.41	409.48 ± 294.12
0.41	411.58 ± 302.37
0.41	412.62 ± 300.41
0.43	405.03 ± 309.95
0.43	405.00 ± 313.06
0.44	404.89 ± 313.29
0.46	413.53 ± 343.50
0.48	422.91 ± 367.87
0.48	423.00 ± 367.90
0.49	432.01 ± 387.04
0.50	439.72 ± 404.35
0.50	454.83 ± 432.82
0.51	454.14 ± 432.60
0.51	457.93 ± 441.05
0.51	461.31 ± 444.62
0.52	468.41 ± 461.85
0.52	475.97 ± 477.10
0.52	479.81 ± 481.34
0.53	460.90 ± 450.01

Continues on next page.

Average $\left(cSpacing \times bps\right)$ continued...

Global G+C	**Average** $\left(cSpacing \times bps\right)$
0.53	484.40 ± 495.35
0.53	500.76 ± 524.25
0.57	573.39 ± 655.29
0.60	671.04 ± 827.91
0.61	693.53 ± 868.95
0.63	771.22 ± 1000.18
0.65	832.41 ± 1102.95
0.65	833.89 ± 1108.87
0.67	936.06 ± 1283.84
0.68	943.01 ± 1301.13
0.68	943.50 ± 1292.37
0.68	950.77 ± 1311.21

Table C.5: This table presents the average $\left(cSpacing \times bps\right)$ values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.9 on page 153.

C.6 Average $(fSpacing \times bps)$ in Random Sequences Across G+C Content Spectrum

GlobalG+C	Average $\left(fSpacing \times bps\right)$
0.25	772.59 ± 627.85
0.26	706.99 ± 555.96
0.29	532.59 ± 428.33
0.29	543.05 ± 433.36
0.29	545.79 ± 428.36
0.31	427.08 ± 355.30
0.31	439.69 ± 356.76
0.32	386.89 ± 314.93
0.33	371.29 ± 307.28
0.33	375.98 ± 306.18
0.33	380.29 ± 318.33
0.35	309.64 ± 258.61
0.35	313.42 ± 261.04
0.36	304.35 ± 255.63
0.36	307.32 ± 259.20
0.37	272.55 ± 237.58
0.37	282.73 ± 236.68
0.38	259.88 ± 222.43
0.38	261.26 ± 222.13
0.38	263.32 ± 227.42
0.38	271.06 ± 231.22
0.38	272.70 ± 233.50
0.39	257.09 ± 221.80
0.40	233.26 ± 209.72
0.40	236.94 ± 208.22
0.40	237.79 ± 210.22
0.41	217.49 ± 190.58

Continues on next page.

Average $\left(fSpacing \times bps\right)$ continued...

Global G+C	**Average** $\left(fSpacing \times bps\right)$
0.41	225.64 ± 194.09
0.41	227.92 ± 201.79
0.41	228.05 ± 197.60
0.43	198.62 ± 177.71
0.43	204.57 ± 182.58
0.44	198.54 ± 177.09
0.46	180.25 ± 164.47
0.48	172.13 ± 160.71
0.48	172.10 ± 159.74
0.49	167.69 ± 155.87
0.50	161.49 ± 151.92
0.50	163.69 ± 154.44
0.51	159.60 ± 150.52
0.51	159.97 ± 152.21
0.51	160.43 ± 151.08
0.52	156.14 ± 147.49
0.52	156.27 ± 148.20
0.52	157.00 ± 149.57
0.53	152.76 ± 144.42
0.53	154.72 ± 144.92
0.53	154.94 ± 150.21
0.57	150.33 ± 145.47
0.60	151.62 ± 146.84
0.61	151.62 ± 146.57
0.63	154.79 ± 151.71
0.65	157.45 ± 152.77
0.65	157.74 ± 153.21
0.67	163.55 ± 159.57
0.68	163.28 ± 158.83

Continues on next page.

Average $\left(fSpacing \times bps\right)$ continued...

Global G+C	**Average $\left(fSpacing \times bps\right)$**
0.68	163.71 ± 159.57
0.68	164.35 ± 160.43

Table C.6: This table presents the average $\left(fSpacing \times bps\right)$ values and their standard deviations measured when the genomes in our test set are randomly shuffled. The search parameters applied by the stem-loop search algorithm to generate these values are cited in Table 3.2 on page 31. The values in this table are graphed in Figure B.11 on page 155.

C.7 *bps* - Z Score Data

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_002162	0.25	10.08	7.25	10.80	10.25
NC_002528	0.26	8.73	5.78	13.63	10.09
NC_001318	0.29	8.05	4.19	11.69	5.47
NC_003366	0.29	6.73	20.3	42.8	20.75
NC_004557	0.29	6.96	21.67	31.19	12.92
NC_000909	0.31	1.41	18.70	35.58	19.16
NC_003030	0.31	-2.33	16.88	37.89	14.95
NC_003103	0.32	16.06	18.65	9.32	8.41
NC_003106	0.33	6.64	8.66	24.63	21.62
NC_003923	0.33	11.71	23.99	26.71	2.55
NC_005791	0.33	7.80	-2.68	25.55	13.86
NC_003997	0.35	14.54	23.40	34.57	15.51
NC_004722	0.35	16.58	27.83	36.38	17.07
NC_003909	0.36	25.45	13.62	35.84	13.82
NC_004368	0.36	2.74	7.31	23.77	8.15
NC_003212	0.37	12.49	11.96	24.94	12.22
NC_004350	0.37	14.78	4.22	21.43	10.38

Continues on next page.

bps Z Scores continued...

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_000907	0.38	16.74	7.35	17.88	9.44
NC_002940	0.38	20.82	9.67	15.81	10.22
NC_003210	0.38	15.02	10.14	23.98	11.86
NC_003869	0.38	2.77	19.77	31.26	14.69
NC_004668	0.38	29.77	21.74	18.94	7.98
NC_002737	0.39	10.29	1.21	17.36	7.05
NC_000912	0.40	27.34	8.60	-0.48	4.64
NC_002620	0.40	-13.97	-3.45	4.63	7.01
NC_002689	0.40	4.82	-11.81	10.24	12.11
NC_000922	0.41	-4.11	-7.53	2.05	3.66
NC_002179	0.41	-3.41	-3.11	5.90	3.02
NC_003901	0.41	73.45	-18.79	19.48	13.08
NC_005043	0.41	-4.68	-6.81	1.83	3.66
NC_000918	0.43	-12.16	-8.01	28.85	13.38
NC_004663	0.43	74.46	13.74	10.34	2.33
NC_000964	0.44	55.52	-5.00	20.31	7.51
NC_000853	0.46	-22.6	-7.54	14.72	10.27
NC_003143	0.48	81.79	-27.97	7.38	1.59
NC_004088	0.48	79.31	-17.58	8.92	-0.98
NC_000917	0.49	-18.87	-8.74	0.90	2.97
NC_000916	0.50	28.28	-25.26	1.84	6.17
NC_004431	0.50	100.67	7.80	-4.28	-5.73
NC_000913	0.51	99.79	-12.77	-5.17	-2.53
NC_002695	0.51	117.23	-11.12	-4.92	-0.34
NC_004741	0.51	84.07	-3.37	-2.07	0.27
NC_003197	0.52	118.64	-16.12	-9.50	-7.05
NC_003198	0.52	117.16	-37.32	-5.49	-4.07
NC_004556	0.52	65.76	-19.96	-7.64	-1.03

Continues on next page.

bps Z Scores continued...

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_000919	0.53	12.04	7.79	-13.10	4.95
NC_002488	0.53	90.18	-1.14	-7.04	-0.38
NC_002935	0.53	54.51	-16.63	-5.74	-0.94
NC_002932	0.57	111.29	-25.06	-15.71	-10.87
NC_004307	0.60	117.93	4.31	-14.40	-12.90
NC_002939	0.61	56.91	-38.97	-26.37	-15.74
NC_004369	0.63	93.03	-36.14	-44.28	-4.90
NC_003919	0.65	246.44	50.87	-25.97	-19.61
NC_005085	0.65	204.43	-16.33	-76.96	-28.30
NC_002696	0.67	174.89	-21.58	-26.60	-12.96
NC_002927	0.68	242.60	-22.81	-48.73	-26.82
NC_002928	0.68	230.51	-19.95	-52.65	-25.53
NC_002929	0.68	216.39	-23.67	-48.60	-18.56

Table C.7: This table presents the Z scores of the average *bps* values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.4 and C.1. The values in this table are graphed in Figure B.2 on page 146.

C.8 *span* - Z Score Data

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002162	0.25	6.12	4.25	10.59	10.28
NC_002528	0.26	6.98	3.27	13.84	9.75
NC_001318	0.29	4.66	3.29	10.90	5.31
NC_003366	0.29	2.70	11.13	40.90	19.69
NC_004557	0.29	2.40	13.57	30.14	12.74
NC_000909	0.31	-1.10	14.76	34.97	18.98
NC_003030	0.31	-4.51	6.81	37.29	15.08
NC_003103	0.32	16.08	15.32	9.39	8.07
NC_003106	0.33	2.72	5.86	24.56	21.25
NC_003923	0.33	5.54	15.83	26.14	2.40
NC_005791	0.33	4.78	-5.28	25.54	13.51
NC_003997	0.35	7.30	5.47	33.72	14.57
NC_004722	0.35	11.67	11.41	35.42	16.09
NC_003909	0.36	19.86	9.31	34.88	13.18
NC_004368	0.36	-3.86	-3.60	22.94	8.43
NC_003212	0.37	13.72	0.89	24.19	11.52
NC_004350	0.37	8.69	-4.12	20.92	9.75
NC_000907	0.38	11.45	0.57	17.90	8.41
NC_002940	0.38	16.46	3.39	15.68	9.26
NC_003210	0.38	15.34	-1.75	23.18	11.14
NC_003869	0.38	-1.19	12.89	30.47	13.96
NC_004668	0.38	22.32	10.96	18.25	6.94
NC_002737	0.39	3.42	-7.01	16.19	7.11
NC_000912	0.40	21.99	6.25	-0.16	3.87
NC_002620	0.40	-21.82	-8.87	4.98	6.50
NC_002689	0.40	4.64	-12.83	10.34	11.08
NC_000922	0.41	-10.95	-12.57	1.70	3.02

Continues on next page.

span Z Scores continued...

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_002179	0.41	-12.38	-8.47	5.47	2.00
NC_003901	0.41	69.56	-24.96	18.64	11.90
NC_005043	0.41	-11.47	-11.94	1.43	3.02
NC_000918	0.43	-17.38	-9.35	28.99	13.16
NC_004663	0.43	67.84	-0.10	9.63	1.36
NC_000964	0.44	54.24	-18.55	19.42	6.56
NC_000853	0.46	-31.35	-11.79	14.25	9.40
NC_003143	0.48	74.06	-36.92	7.32	0.77
NC_004088	0.48	71.42	-25.48	9.01	-1.81
NC_000917	0.49	-22.45	-9.98	0.73	2.54
NC_000916	0.50	21.58	-27.12	1.64	5.56
NC_004431	0.50	90.70	3.49	-4.34	-5.71
NC_000913	0.51	91.07	-22.41	-5.65	-2.27
NC_002695	0.51	108.74	-19.23	-5.03	-0.11
NC_004741	0.51	75.72	-10.24	-2.30	-0.66
NC_003197	0.52	111.59	-23.44	-9.87	-6.70
NC_003198	0.52	110.37	-46.12	-5.71	-4.16
NC_004556	0.52	55.54	-26.56	-7.39	-1.53
NC_000919	0.53	1.43	5.10	-12.27	4.41
NC_002488	0.53	81.04	-6.69	-7.19	-1.18
NC_002935	0.53	43.56	-25.84	-6.42	-1.43
NC_002932	0.57	106.09	-28.09	-16.98	-12.19
NC_004307	0.60	110.41	-1.13	-12.77	-12.58
NC_002939	0.61	50.90	-42.95	-25.76	-16.23
NC_004369	0.63	80.85	-44.37	-45.77	-5.71
NC_003919	0.65	227.81	42.78	-26.57	-19.49
NC_005085	0.65	186.62	-22.61	-78.02	-28.01
NC_002696	0.67	163.57	-27.32	-26.52	-13.15

Continues on next page.

span Z Scores continued...

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002927	0.68	222.73	-29.84	-48.81	-27.36
NC_002928	0.68	212.10	-26.49	-52.26	-25.15
NC_002929	0.68	199.57	-30.36	-48.93	-18.45

Table C.8: This table presents the Z scores of the average *span* values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.5 and C.2. The values in this table are graphed in Figure B.4 on page 148.

C.9 *cSpacing* - Z Score Data

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002162	0.25	-21.40	-0.82	-127.92	-24.15
NC_002528	0.26	-30.75	2.31	-154.55	-7.09
NC_001318	0.29	-16.78	3.66	-82.41	-9.93
NC_003366	0.29	-27.97	17.51	-356.85	-26.16
NC_004557	0.29	-19.70	6.41	-243.29	-21.75
NC_000909	0.31	-4.09	2.07	-154.78	-1.99
NC_003030	0.31	-19.05	23.99	-262.68	-47.24
NC_003103	0.32	-36.16	-3.63	-53.20	-4.87
NC_003106	0.33	-11.90	11.66	-71.83	2.93
NC_003923	0.33	-42.63	16.79	-173.21	-14.55
NC_005791	0.33	-49.83	13.34	-124.20	-1.99
NC_003997	0.35	-42.80	34.25	-167.68	-45.99
NC_004722	0.35	-42.79	29.81	-207.13	-42.07
NC_003909	0.36	-30.98	13.73	-175.02	-40.13
NC_004368	0.36	-36.60	17.02	-116.44	-33.28

Continues on next page.

cSpacing Z Scores continued...

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_003212	0.37	-40.30	29.54	-111.32	-19.71
NC_004350	0.37	-52.40	17.07	-105.26	-26.22
NC_000907	0.38	-77.05	6.71	-62.03	-4.86
NC_002940	0.38	-69.06	-0.03	-59.26	-3.83
NC_003210	0.38	-42.17	31.57	-102.99	-22.20
NC_003869	0.38	-14.44	12.17	-90.41	-23.57
NC_004668	0.38	-68.77	21.08	-65.42	-24.66
NC_002737	0.39	-63.60	12.49	-74.51	-14.55
NC_000912	0.40	-53.54	7.20	-13.98	-21.03
NC_002620	0.40	-45.60	8.02	-28.47	-5.50
NC_002689	0.40	-36.23	19.75	-29.10	1.78
NC_000922	0.41	-40.94	16.42	-11.32	-8.79
NC_002179	0.41	-42.01	16.71	-16.27	-5.18
NC_003901	0.41	-147.31	63.15	-27.43	2.2
NC_005043	0.41	-40.76	17.27	-11.4	-8.49
NC_000918	0.43	-24.56	14.33	-23.54	6.55
NC_004663	0.43	-98.73	58.69	-34.82	-4.31
NC_000964	0.44	-76.94	45.30	-78.82	-19.99
NC_000853	0.46	-58.70	8.10	-17.26	-2.95
NC_003143	0.48	-105.19	37.08	-19.61	-8.49
NC_004088	0.48	-103.62	30.34	-13.89	-10.65
NC_000917	0.49	-16.47	20.85	-1.19	2.18
NC_000916	0.50	-48.86	25.93	-12.08	4.14
NC_004431	0.50	-87.07	8.37	-17.51	-4.22
NC_000913	0.51	-80.13	36.00	-15.52	-2.84
NC_002695	0.51	-76.89	37.58	-15.30	-5.51
NC_004741	0.51	-87.25	13.44	-21.03	-8.29
NC_003197	0.52	-54.47	30.50	-10.30	-9.73

Continues on next page.

cSpacing Z Scores continued...

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_003198	0.52	-55.91	31.13	-22.64	-8.96
NC_004556	0.52	3.33	39.50	-11.53	-0.85
NC_000919	0.53	-23.68	0.46	-7.62	-0.13
NC_002488	0.53	24.41	37.40	-9.59	-2.93
NC_002935	0.53	-24.89	8.43	-7.39	-8.63
NC_002932	0.57	52.10	18.41	-6.20	-6.09
NC_004307	0.60	77.13	12.61	-11.69	-10.76
NC_002939	0.61	3.08	-0.78	-11.09	-9.26
NC_004369	0.63	50.16	-14.94	-28.70	-5.01
NC_003919	0.65	197.38	47.15	-18.24	-16.47
NC_005085	0.65	164.17	10.20	-29.02	-13.18
NC_002696	0.67	143.59	-17.71	-20.53	-12.47
NC_002927	0.68	209.64	-4.16	-28.01	-18.80
NC_002928	0.68	200.36	-2.49	-25.73	-17.49
NC_002929	0.68	190.20	-2.40	-28.18	-12.99

Table C.9: This table presents the Z scores of the average *cSpacing* values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.6 and C.3. The values in this table are graphed in Figure B.6 on page 150.

C.10 *fSpacing* - Z Score Data

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002162	0.25	-21.56	-1.01	-133.91	-26.13
NC_002528	0.26	-31.01	1.89	-190.08	-8.13
NC_001318	0.29	-16.85	3.46	-90.81	-11.24
NC_003366	0.29	-27.84	16.99	-465.18	-29.03
NC_004557	0.29	-19.60	5.63	-297.31	-24.17
NC_000909	0.31	-3.55	0.74	-495.98	-4.83
NC_003030	0.31	-18.23	23.70	-350.26	-52.17
NC_003103	0.32	-38.11	-5.69	-66.22	-6.45
NC_003106	0.33	-11.53	11.21	-156.33	0.48
NC_003923	0.33	-43.15	15.09	-220.26	-29.67
NC_005791	0.33	-49.60	14.02	-226.31	-4.08
NC_003997	0.35	-43.32	33.90	-244.91	-61.19
NC_004722	0.35	-44.12	28.67	-320.36	-52.12
NC_003909	0.36	-34.77	12.21	-259.61	-50.43
NC_004368	0.36	-35.30	17.61	-156.80	-38.09
NC_003212	0.37	-43.15	29.62	-171.55	-26.68
NC_004350	0.37	-53.00	18.10	-152.85	-32.12
NC_000907	0.38	-79.68	7.10	-84.34	-7.08
NC_002940	0.38	-72.97	-0.04	-74.31	-6.42
NC_003210	0.38	-45.45	32.41	-159.45	-30.35
NC_003869	0.38	-12.08	10.68	-200.01	-32.20
NC_004668	0.38	-72.35	19.77	-92.40	-28.17
NC_002737	0.39	-63.81	14.22	-94.69	-18.00
NC_000912	0.40	-62.02	6.01	-14.04	-26.12
NC_002620	0.40	-36.82	10.69	-35.35	-7.62
NC_002689	0.40	-36.56	22.08	-56.07	-0.77
NC_000922	0.41	-36.18	19.08	-12.36	-9.78

Continues on next page.

fSpacing Z Scores continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	**Z Score**	**Z Score**	**Z Score**	**Z Score**
NC_002179	0.41	-37.10	18.35	-21.83	-5.53
NC_003901	0.41	-176.98	69.07	-47.36	-0.57
NC_005043	0.41	-35.80	19.72	-12.23	-9.32
NC_000918	0.43	-14.05	17.72	-165.88	2.56
NC_004663	0.43	-125.11	60.37	-43.58	-4.52
NC_000964	0.44	-99.69	51.20	-124.35	-27.15
NC_000853	0.46	-38.67	13.81	-68.98	-8.93
NC_003143	0.48	-169.60	54.75	-29.94	-9.50
NC_004088	0.48	-165.71	44.84	-23.04	-9.20
NC_000917	0.49	2.80	28.10	-3.47	2.21
NC_000916	0.50	-63.05	38.08	-17.04	3.23
NC_004431	0.50	-183.45	15.07	-15.78	-0.59
NC_000913	0.51	-187.85	56.65	-13.5	-0.85
NC_002695	0.51	-193.40	57.12	-13.14	-5.00
NC_004741	0.51	-180.58	31.83	-24.70	-7.19
NC_003197	0.52	-187.86	56.79	-4.13	-4.37
NC_003198	0.52	-183.72	64.01	-25.21	-5.92
NC_004556	0.52	-42.96	59.20	-7.32	1.36
NC_000919	0.53	-12.79	0.46	0.37	-3.50
NC_002488	0.53	-47.06	47.27	-4.83	-0.95
NC_002935	0.53	-75.49	35.42	-1.18	-10.71
NC_002932	0.57	-67.34	53.18	6.67	3.32
NC_004307	0.60	-59.66	34.90	2.37	0.11
NC_002939	0.61	-49.50	65.03	10.59	4.47
NC_004369	0.63	-35.10	55.55	13.31	0.60
NC_003919	0.65	-50.15	38.63	6.43	-0.46
NC_005085	0.65	-5.02	79.73	26.71	9.39
NC_002696	0.67	-21.14	41.27	9.05	-1.27

Continues on next page.

fSpacing Z Scores continued...

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002927	0.68	15.29	75.42	14.12	7.81
NC_002928	0.68	19.49	72.16	15.24	8.31
NC_002929	0.68	28.47	75.98	13.67	7.97

Table C.10: This table presents the *Z* scores of the average *fSpacing* values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.7 and C.4. The values in this table are graphed in Figure B.8 on page 152.

C.11 (*cSpacing* × *bps*) - Z Score Data

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002162	0.25	-16.98	2.96	-57.76	-3.88
NC_002528	0.26	-24.27	4.25	-28.35	2.63
NC_001318	0.29	-11.40	4.66	-27.66	-0.44
NC_003366	0.29	-21.37	23.19	-61.66	-2.31
NC_004557	0.29	-16.30	15.78	-39.16	-3.07
NC_000909	0.31	-5.57	10.37	4.61	9.54
NC_003030	0.31	-19.13	25.82	-45.89	-10.82
NC_003103	0.32	-19.39	8.99	-6.80	3.36
NC_003106	0.33	-9.57	11.71	6.42	12.74
NC_003923	0.33	-28.87	24.34	-32.07	-0.98
NC_005791	0.33	-35.04	8.10	-3.16	7.07
NC_003997	0.35	-24.41	35.58	-19.58	-2.98
NC_004722	0.35	-11.04	35.24	-24.83	-3.98
NC_003909	0.36	7.64	13.26	-19.28	-3.94
NC_004368	0.36	-26.80	15.51	-13.44	-6.78

Continues on next page.

$\left(cSpacing \times bps\right)$ Z Scores continued...

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_003212	0.37	-19.36	25.28	-8.68	1.83
NC_004350	0.37	-27.23	13.47	-10.81	-3.57
NC_000907	0.38	-33.68	8.16	-5.96	4.41
NC_002940	0.38	-24.52	5.56	-10.56	5.09
NC_003210	0.38	-17.02	24.98	-7.47	1.57
NC_003869	0.38	-10.06	18.16	6.16	1.71
NC_004668	0.38	-24.29	25.81	-5.22	-4.42
NC_002737	0.39	-34.85	7.73	-8.83	-0.82
NC_000912	0.40	-3.08	8.82	-8.08	-3.60
NC_002620	0.40	-37.97	2.22	-7.20	2.99
NC_002689	0.40	-18.24	5.19	1.77	8.28
NC_000922	0.41	-26.65	5.43	-3.24	-0.56
NC_002179	0.41	-24.65	7.80	-0.68	-0.03
NC_003901	0.41	-12.00	30.89	6.66	9.59
NC_005043	0.41	-27.26	6.56	-3.59	-0.47
NC_000918	0.43	-21.55	2.22	16.33	11.68
NC_004663	0.43	11.05	37.50	-4.24	0.05
NC_000964	0.44	8.35	20.81	-5.15	-1.75
NC_000853	0.46	-40.47	-2.67	7.60	6.68
NC_003143	0.48	33.51	-3.53	-0.66	-1.09
NC_004088	0.48	32.21	-0.44	1.76	-3.37
NC_000917	0.49	-19.22	1.02	0.60	2.64
NC_000916	0.50	8.80	-5.24	-1.92	5.93
NC_004431	0.50	57.38	9.29	-10.27	-6.71
NC_000913	0.51	58.82	4.36	-10.78	-2.00
NC_002695	0.51	71.87	5.16	-11.40	-1.23
NC_004741	0.51	46.26	2.43	-7.50	-1.47
NC_003197	0.52	80.00	-0.02	-14.54	-9.57

Continues on next page.

$\left(cSpacing \times bps\right)$ Z Scores continued...

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_003198	0.52	78.53	-16.91	-12.00	-6.10
NC_004556	0.52	52.65	0.26	-9.37	-0.62
NC_000919	0.53	7.56	6.85	-16.58	4.10
NC_002488	0.53	67.63	12.11	-8.26	-0.78
NC_002935	0.53	42.12	-8.99	-6.49	-2.48
NC_002932	0.57	92.89	-10.65	-16.60	-9.96
NC_004307	0.60	101.14	11.80	-14.20	-11.96
NC_002939	0.61	54.08	-22.70	-26.02	-15.13
NC_004369	0.63	85.71	-26.71	-45.24	-5.02
NC_003919	0.65	210.19	50.30	-27.24	-21.26
NC_005085	0.65	171.46	-0.25	-86.68	-30.70
NC_002696	0.67	155.78	-9.66	-27.22	-11.95
NC_002927	0.68	211.95	-8.49	-57.64	-28.92
NC_002928	0.68	199.95	-7.37	-61.78	-27.82
NC_002929	0.68	190.15	-11.87	-57.62	-18.18

Table C.11: This table presents the Z scores of the average $\left(cSpacing \times bps\right)$ values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.8 and C.5. The values in this table are graphed in Figure B.10 on page 154.

C.12 ($fSpacing \times bps$) - Z Score Data

Accession Number	Global G+C	CDS Z Score	NC Z Score	rRNA Z Score	tRNA Z Score
NC_002162	0.25	-16.98	2.96	-57.76	-3.88
NC_002528	0.26	-24.27	4.25	-28.35	2.63
NC_001318	0.29	-11.40	4.66	-27.66	-0.44
NC_003366	0.29	-21.37	23.19	-61.66	-2.31
NC_004557	0.29	-16.30	15.78	-39.16	-3.07
NC_000909	0.31	-5.57	10.37	4.61	9.54
NC_003030	0.31	-19.13	25.82	-45.89	-10.82
NC_003103	0.32	-19.39	8.99	-6.80	3.36
NC_003106	0.33	-9.57	11.71	6.42	12.74
NC_003923	0.33	-28.87	24.34	-32.07	-0.98
NC_005791	0.33	-35.04	8.10	-3.16	7.07
NC_003997	0.35	-24.41	35.58	-19.58	-2.98
NC_004722	0.35	-11.04	35.24	-24.83	-3.98
NC_003909	0.36	7.64	13.26	-19.28	-3.94
NC_004368	0.36	-26.80	15.51	-13.44	-6.78
NC_003212	0.37	-19.36	25.28	-8.68	1.83
NC_004350	0.37	-27.23	13.47	-10.81	-3.57
NC_000907	0.38	-33.68	8.16	-5.96	4.41
NC_002940	0.38	-24.52	5.56	-10.56	5.09
NC_003210	0.38	-17.02	24.98	-7.47	1.57
NC_003869	0.38	-10.06	18.16	6.16	1.71
NC_004668	0.38	-24.29	25.81	-5.22	-4.42
NC_002737	0.39	-34.85	7.73	-8.83	-0.82
NC_000912	0.40	-3.08	8.82	-8.08	-3.60
NC_002620	0.40	-37.97	2.22	-7.20	2.99
NC_002689	0.40	-18.24	5.19	1.77	8.28
NC_000922	0.41	-26.65	5.43	-3.24	-0.56

Continues on next page.

$\left(fSpacing \times bps\right)$ Z Scores continued...

Accession	**Global**	**CDS**	**NC**	**rRNA**	**tRNA**
Number	**G+C**	**Z Score**	**Z Score**	**Z Score**	**Z Score**
NC_002179	0.41	-24.65	7.80	-0.68	-0.03
NC_003901	0.41	-12.00	30.89	6.66	9.59
NC_005043	0.41	-27.26	6.56	-3.59	-0.47
NC_000918	0.43	-21.55	2.22	16.33	11.68
NC_004663	0.43	11.05	37.50	-4.24	0.05
NC_000964	0.44	8.35	20.81	-5.15	-1.75
NC_000853	0.46	-40.47	-2.67	7.60	6.68
NC_003143	0.48	33.51	-3.53	-0.66	-1.09
NC_004088	0.48	32.21	-0.44	1.76	-3.37
NC_000917	0.49	-19.22	1.02	0.60	2.64
NC_000916	0.50	8.80	-5.24	-1.92	5.93
NC_004431	0.50	57.38	9.29	-10.27	-6.71
NC_000913	0.51	58.82	4.36	-10.78	-2.00
NC_002695	0.51	71.87	5.16	-11.40	-1.23
NC_004741	0.51	46.26	2.43	-7.50	-1.47
NC_003197	0.52	80.00	-0.02	-14.54	-9.57
NC_003198	0.52	78.53	-16.91	-12.00	-6.10
NC_004556	0.52	52.65	0.26	-9.37	-0.62
NC_000919	0.53	7.56	6.85	-16.58	4.10
NC_002488	0.53	67.63	12.11	-8.26	-0.78
NC_002935	0.53	42.12	-8.99	-6.49	-2.48
NC_002932	0.57	92.89	-10.65	-16.60	-9.96
NC_004307	0.60	101.14	11.80	-14.20	-11.96
NC_002939	0.61	54.08	-22.70	-26.02	-15.13
NC_004369	0.63	85.71	-26.71	-45.24	-5.02
NC_003919	0.65	210.19	50.30	-27.24	-21.26
NC_005085	0.65	171.46	-0.25	-86.68	-30.70
NC_002696	0.67	155.78	-9.66	-27.22	-11.95

Continues on next page.

$\left(fSpacing \times bps\right)$ Z Scores continued...

Accession Number	**Global G+C**	**CDS Z Score**	**NC Z Score**	**rRNA Z Score**	**tRNA Z Score**
NC_002927	0.68	211.95	-8.49	-57.64	-28.92
NC_002928	0.68	199.95	-7.37	-61.78	-27.82
NC_002929	0.68	190.15	-11.87	-57.62	-18.18

Table C.12: This table presents the Z scores of the average $\left(fSpacing \times bps\right)$ values in the various genomic domains of the genomes in our test set. These values were calculated using the tables in Appendices A.9 and C.6. The values in this table are graphed in Figure B.12 on page 156.